Capítulo 14

Estructura, plegamiento y evolución del RNA

Ivan Dotu, Michael Stich y Jacobo Aguirre

14.1. Introducción

La estructura de ácidos nucleicos más famosa es el DNA cromosomal, donde el DNA se encuentra en forma de dos cadenas enlazadas – la célebre doble hélice. Esta estructura formada por pares de bases complementarias permite relativamente pocas interacciones físico-químicas con otras moléculas. Ese no es el caso del RNA que en muchos contextos funcionales se encuentra en forma de cadena sencilla. Por consecuencia, puede dar lugar a la formación de enlaces con otras moléculas o permitir interacciones secundarias y terciarias consigo mismo que representan el *plegamiento espacial* de la molécula. Este plegamiento es responsable de la versatilidad funcional biológica del RNA.

El RNA, por su composición química, es muy similar al DNA y puede ejercer algunas funciones similares. No obstante, mientras el papel del DNA en los organismos actuales es –en primera aproximación– limitado a constituir el genoma, el RNA también puede ejercer un papel catalítico o regulador, como las proteínas. En este sentido *proteínas y RNA son similares a nivel estructural-funcional.* Como *el RNA es la única biomolécula que a la vez puede codificar su genoma y ejercer un papel catalítico*, se ha convertido en una molécula clave para entender la evolución temprana de la vida. En particular, se ha desarrollado la hipótesis de "Mundo RNA" en la que se supone que la vida actual fue precedida por una etapa donde el RNA realizaba tanto las funciones del DNA como las de las proteínas.

Resumiendo, la relevancia del RNA reside no sólo en codificar la información que finalmente puede ser traducida a proteínas, sino en regular y catalizar procesos celulares a través de su conformación espacial, es decir, su estructura.

14.2. Tipos de RNA

El RNA aparece en muchos procesos celulares. Quizá el más conocido sea el *RNA mensajero*, mRNA. En este caso, el RNA representa una parte de la información genética que codifica para una o varias proteínas. No obstante, el avance científico permite identificar y entender cada vez más procesos en los

que el papel del RNA es *no-codificante*, y la clasificación y nomenclatura de los tipos de RNA está cambiando constantemente. Un caso particular representan los virus que utilizan RNA para codificar su genoma, los llamados *virus RNA*. En la Tabla 14.1 presentamos una lista no-exhaustiva de tipos de RNA.

Nombre	Abreviatura	Función
RNA en síntesis de proteínas		
RNA mensajero	mRNA	Codificante de proteínas
RNA ribosomal	rRNA	Traducción
RNA de transferencia	tRNA	Traducción
RNA reguladores		
RNA antisentido	aRNA	Regulación de mRNA
RNA largo no codificante	long ncRNA	Varias
Micro RNA	miRNA	Regulación de genes
Small interfering RNA	siRNA	Regulación de genes
Piwi-interacting RNA	piRNA	Defensa contra transposones
Riboswitch		Regulación de un gen en el mRNA
Ribozima		Catálisis de reacciones químicas
RNA en modificación postranscripcional o replicación de DNA		
Espliceosoma	snRNA	Splicing
RNA pequeño nucleolar	snoRNA	Modificación de nucleótidos
Ribonucleasa P	RNase P	Maduración tRNA
Ribonucleasa MRP	RNase MRP	Maduración rRNA, replicación DNA
RNA telomerasa	telRNA	Síntesis de DNA telomérico
RNA "parásito"		
Virus de RNA		Codificante
Viroides		Auto-propagante
RNA satélite		Auto-propagante
Retrotransposón		Auto-propagante

Tabla 14.1: Tipos de RNA. Los tipos miRNA, siRNA y piRNA se pueden agrupar como RNA de interferencia, iRNA. El concepto de RNA no-codificante, ncRNA, agrupa los RNA reguladoras y los RNA en modificación postranscripcional o replicación de DNA. También existe el término fRNA, RNA funcional, que frequentemente es utilizado como sinónimo a ncRNA.

Describimos brevemente algunos tipos de RNA relevantes. El *RNA mensajero* es el RNA codificante por excelencia. Es la transcripción del DNA que posteriormente es traducido en proteínas. El mRNA está caracterizado por su secuencia. Pero en realidad, el mRNA puede plegarse y formar estructuras secundarias que, hasta la fecha, no creemos que tengan función. Una vez en el citoplasma, la propia maquinaria de traducción tiene que incluir un mecanismo para deshacer estas estructuras.

En el proceso de traducción, el mRNA está introducido en el ribosoma. El ribosoma es un complejo macromolecular formado por varias partes y constituido aproximadamente por un tercio de proteínas y dos tercios de RNA, el *RNA ribosomal* (o ribosómico). Ese rRNA es central al proceso de traducción, y la formación de enlaces peptídicos es catalizado por rRNA. El rRNA consiste de varios miles de nucleótidos repartidos en 3 (procariotas) o 4 (eucariotas) fragmentos. El rRNA se encuentra plegado

tridimensionalmente.

El *RNA de transferencia*, tRNA, actúa conjuntamente con el ribosoma y se encarga de traducir un triplete del mRNA en su amino ácido correspondiente. Su longitud típica es de alrededor de 76 nt. Mientras la secuencia que forma el tRNA puede variar entre especies distintas, la molécula tiene una estructura secundaria (y terciaria) similar a una hoja de trébol con cuatro brazos principales, muy conservada entre especies, lo que demuestra que es la estructura espacial de la molécula la que confiere su función.

Como la molécula de RNA puede formar estructuras secundarias y terciarias complejas, similar a las proteínas, también les da la capacidad de catalizar reacciones químicas, al igual que las enzimas. Las moléculas de RNA cuya función principalmente es esa se llaman *ribozimas*, como el *hairpin ribozyme* o el *hammerhead ribozyme*. Son moléculas de varias decenas de nucleótidos.

Por último, los *riboswitches* son partes del mRNA localizadas antes (5'-UTR) o después (3'-UTR) de un gen. Un *riboswitch* cambia su conformación espacial (su forma plegada) si se le une una molecula especifica y de ese modo regula la traducción y expresión del gen correspondiente.

14.3. Niveles estructurales del RNA

En RNA, tenemos tres tipos de niveles estructurales: la estructura primaria, secundaria y terciaria. Dicha jerarquía estructural es análoga a la descrita para proteínas (Sección 7.6)

La *estructura primaria* de una molécula de RNA está descrita por su secuencia, 5'-GAACGUUG...-3' (ver Sección 7.3). La secuencia es una entidad lineal, con una longitud dada por el número de nucleótidos que la forman y tiene un comienzo (5') y un final (3'). Un RNA de cadena sencilla sin plegar no sería más que una hélice abierta (Figura 14.2). Esa situación es altamente inestable porque las bases de una parte de la molécula pueden formar enlaces con bases de otra parte. Este proceso se llama *plegamiento* y da lugar a la estructura secundaria y estructura terciaria.

La *estructura secundaria* es un estado intermedio del plegamiento y caracterizado por la formación de pares de bases (sobre todo pares del tipo Watson-Crick) y bucles. Una estructura secundaria puede ser descrita en un plano (bidimensional).

La *estructura terciaria* representa el estado final del plegamiento que incluye posibles interacciones entre partes cercanas y/o lejanas de la molécula. La formación de enlaces va más allá de simples pares tipo Watson-Crick y la descripción de la molécula es tridimensional.

La *estructura secundaria* puede estudiarse experimentalmente utilizando métodos enzimáticos o de modificación química, que utilizan compuestos químicos como DMS, ketoxal o NMIA, que reaccionan específicamente sobre los nucleótidos desapareados y la estructura terciaria a través de cristalografía de rayos X o de resonancia magnética de la sustancia cristalizada (ver sec:Macromoleculas:MetodosExperimentalesEstruc
En la Sección 14.4 veremos el proceso de plegamiento en más detalle.

14.3.1. Composición química y estructura primaria

El ácido ribonucléico (RNA, por su siglas en inglés, *ribonucleic acid*) es una de las biomoléculas más importantes. Su composición es similar al ácido desoxirribonucléico (DNA) y su amplio abanico de funciones incluye propiedades del DNA y de las proteínas, lo que convierte a esta molécula en una de las biomoléculas más versátiles. Para distinguir moléculas de cadena simple y doble (sobre todo

en el contexto de virus de RNA), se usa la nomenclatura ssRNA (*single-stranded* RNA) y dsRNA (*double-stranded* RNA).

En el caso más simple y relevante, el RNA está formado por una cadena sencilla de azúcar (*ribosa*) y un *grupo fosfato* a la que se unen cuatro tipos de *bases nitrogenadas*, adenina, citosina, guanina y uracilo, abreviados por las primeras letras de sus nombres, A, C, G y U. Mientras A y G son purinas (con una estructura de dos anillos heterocíclicos fusionados), C y U son pirimidinas (con un anillo heterocíclico) (ver Sección 7.3). En la naturaleza existen de forma minoritaria (pero en moléculas importantes como el tRNA) *modificaciones de las bases canónicas* de las cuales mencionamos algunas: pseudouridina (Ψ o P), dihidrouridina (D), inosina (I), 7-metilguanosina (m7G o 7). Para ver una lista más completa con sus abreviaturas [65] y para otras notaciones (p. ej. Y pirimidina, R purina), ver las recomendaciones de la International Union of Biochemistry[1].

La unión formada por la base y la ribosa se llama *nucleósido*, y si añadimos además el grupo fosfato, se llama *nucleótido*. Una molécula de RNA está completamente caracterizada por la secuencia de las bases. La secuencia de la molécula coincide con su *estructura primaria*.

Por el posicionamiento relativo de base, ribosa y grupo fosfato, una secuencia no es idéntica a su secuencia inversa. Por lo tanto, una cadena sencilla de RNA tiene un comienzo, denominado 5', y un final, 3'. En la Figura 14.1 se muestra de forma esquemática la secuencia 5'-CGAU-3', la posición de la ribosa, del grupo fosfato y de la base en el RNA. La unión entre grupo fosfato y la ribosa es llamado *esqueleto azúcar-fosfato*. Los nucleótidos forman una cadena al tener enlaces fosfodiéster entre ellos. Cada grupo fosfato tiene una carga negativa neta. Para estabilizar la cadena, el medio suele tener iones positivos (p. ej. potasio, magnesio, sodio).

El RNA, igual que el DNA, es un *polinucleótido*, y puede ser interpretado como un polímero lineal y aperiódico, cuyos elementos, los monómeros, son los nucleótidos. El enlace covalente que proporciona la conexión entre los monómeros es un *enlace fosfodiéster* que forma una unión éster entre el grupo OH del carbono 3' de la ribosa en el monómero anterior con el ácido fosfórico y otra unión éster entre el ácido fosfórico y el carbono 5' de la ribosa en el monómero posterior, resultando en el enlace fosfodiéster 3'-5' que incorpora todo el grupo fosfato. Entre un enlace y el siguiente se forma un ángulo y en consecuencia una cadena simple de RNA gira alrededor de su eje central y forma una *hélice dextrógira* (como se muestra en la Figura 14.2). Si además el RNA forma pares de bases (véase Sección 14.3), la hélice resultante se llama A-RNA (RNA-11), que tiene 11 nucleótidos para formar una giro entero.

Hay dos diferencias fundamentales entre RNA y DNA a nivel de la composición (Figura 14.2). Mientras en el RNA el azúcar es la ribosa, en el DNA es la desoxirribosa, que se diferencia de la ribosa en tener un H en la posición 2' de la pentosa, donde el RNA tiene un OH, lo que confiere una mayor estabilidad al DNA en comparación al RNA. La segunda diferencia importante es la base uracilo que en el DNA está reemplazado por la timina (también una pirimidina). Además, la doble hélice estándar de RNA se encuentra en forma A, la del DNA en forma B.

14.3.2. Estructura secundaria

Las bases de una secuencia de RNA pueden formar *enlaces de hidrógeno* entre sí, formando un par de bases, siendo los más importantes los pares de bases tipo Watson-Crick (o *canónicos*), G-C, C-G, A-U y U-A, pero también los pares *wobble* G-U y U-G. En la Figura 14.3(a) se ve que el par G-C tiene tres enlaces de hidrógeno, y los pares A-U y G-U dos. Dadas esas posibilidades de emparejamiento,

[1] Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. `http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html`

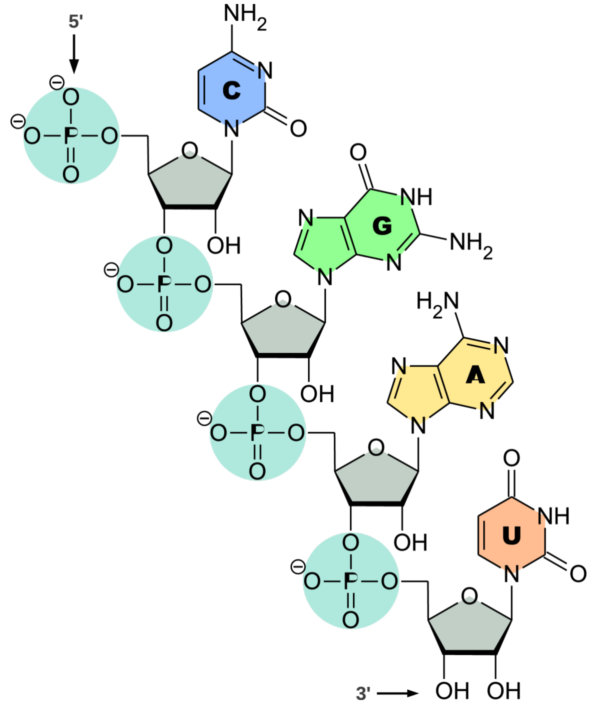

Figura 14.1: Composición química del RNA. Se muestra de forma esquemática la secuencia 5'-CGAU-3', la posición de la ribosa (en gris), el grupo fosfáto (azul) y de la base en el RNA (colores distintos según la base). Se distinguen claramente las purinas (dos anillos) de las pirimidinas (un anillo). El enlace entre nucleótidos es un enlace fosfodiéster. Las cargas negativas están compensadas por iones positivos en el medio (p. ej. K^+, Mg^{2+}, Na^+).

dos nucleótidos elegidos al azar forman un par (son compatibles) con una probabilidad del 37.5 %. No obstante, en secuencias reales –y por lo tanto no aleatorias– el número de bases que forman pares en la estructura secundaria suele ser más alto, p. ej., en un tRNA, alrededor del 60 %. La formación de pares de bases está esquemáticamente dibujado en la Figura 14.3(b). En general, se forman zonas con varios pares de bases seguidos, formando así un *apilamento* o *stack*. La longitud de un *stack* se mide en número de pares de bases consecutivos, bp, por su siglas en inglés *base pairs*. Si un *stack* es suficientemente largo, forma localmente un estructura de doble hélice, también llamado *dúplex*.

Una observación fundamental en el plegamiento del RNA es que si la molécula forma enlaces consigo mismo, necesariamente tiene que formar por lo menos un bucle para acercar los nucleótidos suficientemente (véase la Figura 14.3(b)). Ese bucle se llama bucle terminal aunque es más común llamarlo *bucle horquilla* o *hairpin loop*. Una molécula necesita espacio para realizar este giro, lo que implica que hay por lo menos 2 o 3 nucleótidos dentro del bucle que no pueden participar en la formación de un par de bases. Este argumento también prohíbe la formación de pares de bases entre bases vecinas. En una estructura secundaria un nucleótido o está sin aparear o forma parte de un sólo par. Las bases sin aparear que no se encuentran en un bucle, son llamadas exteriores.

Para describir una estructura secundaria, se suele utilizar puntos y paréntesis. Un punto "." refleja una base sin aparear, un paréntesis "(" una base apareada con una base hacía el 3' de la secuencia, y un paréntesis ")" una base con una pareja hacía el 5' (existen otras notaciones, pero aquí nos limitamos a esta). El número de paréntesis abiertos y cerrados tiene que ser idéntico y obviamente sólo pueden cerrarse paréntesis que hayan sido abiertos anteriormente. A muchos efectos, además, se considera una estructura secundaria ser libre de *pseudonudos* (*pseudoknots*), es decir, pares entrelazados están

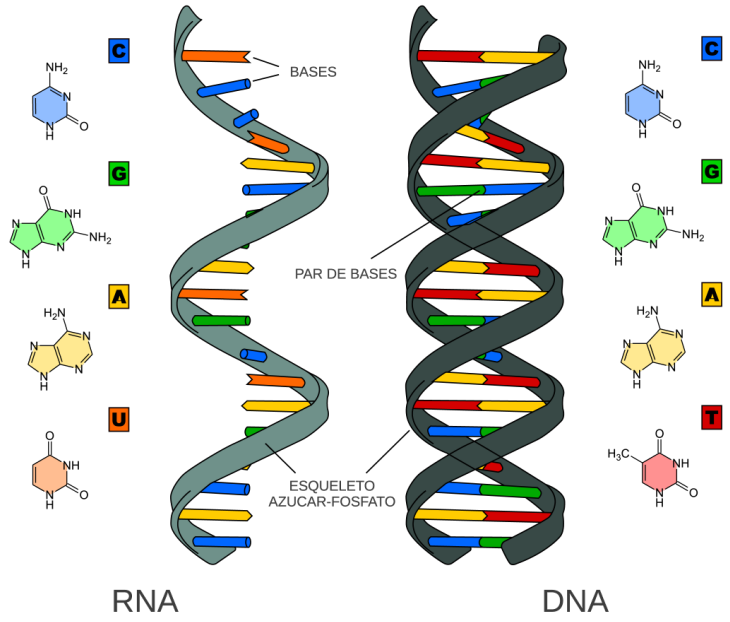

Figura 14.2: RNA y DNA. El RNA (izquierda) tiene la base uracilo en vez de la base timina del DNA (derecha). El RNA contiene ribosa, el DNA desoxirribosa y por lo tanto el esqueleto azúcar-fosfato tiene propiedades físico-químicas distintas. Mientras el DNA típicamente está presente como cadena doble, el RNA suele aparecer como cadena sencilla. La imagen es esquemática ya que en realidad existen un surco mayor y menor en la doble hélice. Fuente: Wikipedia.

prohibidos (las líneas en la Figura 14.3(b) no se pueden cruzar.)

En la Figura 14.3(c), presentamos los elementos fundamentales de estructuras secundarias. Además del bucle terminal, limitado por un sólo par de bases, también existen *bulges* y bucles interiores, flanqueados por pares de bases distintas. Un bucle donde confluyen más de dos *stacks*, se llama *bucle múltiple*. Existen varios esquemas para clasificar estructuras secundarias en términos de esos elementos estructurales. Es obvio que la estructura secundaria más simple está formada por un *stack* y un bucle terminal. Esa estructura se llama *stem-loop* aunque también es llamada *hairpin* en ciertos contextos.

14.3.3. Estructura terciaria

La formación local de pequeñas hélices y bucles es central en el plegamiento del RNA, pero en muchos casos, sobre todo para moléculas grandes, no describe la estructura definitiva tridimensional (terciaria) satisfactoriamente. Describimos tres tipos de interacciones que van más allá de la estructura secundaria: formación de pares de bases no-canónicas, apilamiento de hélices, y pseudonudos.

Analizando estructuras reales, se ha encontrado muchos pares de bases que no son pares canónicos (o *wobble*, en este contexto): ej. en rRNA sólo dos tercios de todos los pares de bases reales son pares canónicos y *wobble* [70]. También se puede dar la situación que lo que parece un bucle interior, en realidad es una hélice con pares no-canónicas (ej. el *E loop*). Para explicar estos resultados tenemos que consider la extensión tridimensional y la posición real (Figura 14.1 y Figura 14.2) de todos los componentes que forman el RNA, es decir, el grupo fosfato, la ribosa y la base nitrogenada. En general, un nucleótido tiene tres lados de interacción: El lado "Watson-Crick" (WC) es el lado estándar, pero también se observan enlaces vía el lado "Hoogsteen" o el lado de la ribosa (*sugar edge* en inglés).

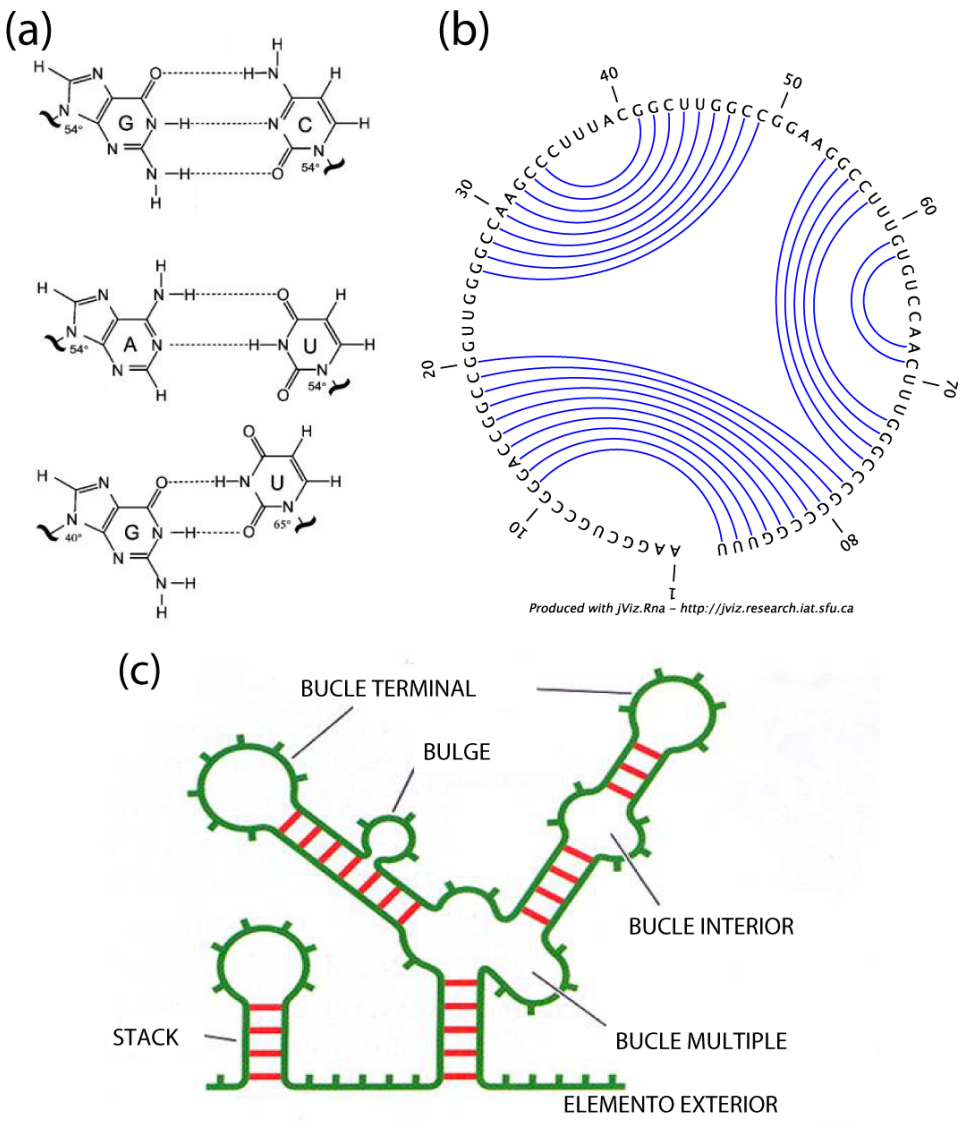

Figura 14.3: Estructuras secundarias. (a) Pares de bases Watson-Crick G-C (arriba), Watson-Crick A-U (centro) y *wobble* G-U (abajo). Figura basada en la Fig. 1 de [74]. (b) Representación circular de la estructura secundaria. Las líneas curvadas indican pares de bases. (c) Elementos de estructuras secundarias. Los pares de bases (lineas rojas) se pueden apilar y forman *stacks*. Las bases sin aparear aparecen como parte de elementos externos o dentro de bucles. Entre los bucles se distinguen los bucles terminales (en inglés, *hairpin*), con un par que cierra el bucle, los *bulges* y bucles interiores, que tienen dos pares que cierran el bucle, y los bucles múltiples, que tienen más que dos pares que cierran el bucle. *Bulges* son bucles con bases sin aparear en un solo lado de la cadena, bucles interiores tienen bases sin aparear en los dos lados.

Cada uno de los seis posibles tipos de enlace entre dos nucleótidos existe en dos variantes, según la orientación (*cis* o *trans*) de los enlaces entre base y azúcar. Como resultado hay 12 combinaciones posibles de formación de pares de bases entre dos nucleótidos en estructuras terciarias. Los pares canónicos G-C, C-G, A-U, U-A, y *wobble* todos pertenecen a un sólo grupo: WC-WC en cis. Pero existen datos experimentales de todas las clases [70]. Además, también se dan enlaces de hidrógeno entre 3 o 4 nucleótidos a la vez, llamados triplexes y cuadruplexes.

Otro efecto observado en estructuras terciarias es el apilamiento coaxial de diferentes hélices, en inglés *coaxial* o *helical stacking*. Es energéticamente favorable que diferentes hélices se colocan de forma paralela y apilada a pesar de que eso aparentamente distorsiona la estructura secundaria. Más abajo veremos un ejemplo en el tRNA.

Los *pseudonudos* son otro elemento importante de estructuras terciarias: Se dan cuando nucleótidos sin aparear en la estructura secundaria forman pares entrelazandose con hélices de la estructura secundaria, violando la condición dada arriba. En ese caso, las líneas que describen los pares se cruzan en la representación circular de estructuras secundarias (Figura 14.3(b)).

Otros motivos estructurales terciarios importantes son *kissing hairpins* (las bases de dos bucles terminales distintos forman pares), el motivo *A-minor*, el receptor tetrabucle, o el *ribose zipper*. Para más información, ver [10, 59].

14.3.4. Ejemplo estructura tRNA

El tRNA es una molécula central para la traducción de la información genética, con una estructura muy conservada entre todos los organismos vivos. En la Figura 14.4(a) vemos una secuencia de tRNA. Además de las bases canónicas, vemos algunas bases modificadas, típico para el tRNA. En (b) está la estructura secundaria en notación punto-paréntesis. En (c) mostramos la estructura secundaria del RNA como imagen planar: las líneas cortas indican un par de base Watson-Crick o *wobble*. La estructura se conoce como hoja de trébol y está formada por cuatro *stacks*, tres bucles terminales y un bucle múltiple. Pero de hecho, el diagrama incluye ya interacciones terciarias (indicadas por las líneas largas). Para apreciar mejor la configuración espacial de la estructura, la dibujamos en (d) desde el lado. Las líneas largas de color indican por donde sigue la secuencia, las líneas negras indican interacciones terciarias. Finalmente, en (e) la representación tridimensional del tRNA basada en datos experimentales con su forma real de "L". Tanto las hélices T y *Acceptor*, como D y *Anticodon* están apiladas, siendo ejemplos de apilamiento coaxial. También existe una interacción bucle-bucle de los nucleótidos 18 y 19 con 55 y 56, y una interacción terciaria AT en el bucle T. Mientras la estructura secundaria considera como apareados sólo 41 de 76 nt, de hecho 72 nt participan en interacciones terciarias (incluyendo apilamiento coaxial).

14.3.5. Arquitectura del RNA

Métodos modernos ofrecen una vista mucho más completa y detallada de moléculas de RNA largas, de complejos RNA-proteínas, virus RNA, etc. Sin renunciar a la descripción clásica como estructura secundaria o terciaria, es útil describir y clasificar moléculas largas a través de los motivos estructurales que la forman, donde un motivo puede ser una simple hélice o elementos más complejos como *ribose zipper* o hélices apiladas. Esa clasificación en motivos estructurales o funcionales (muchas veces recurrentes) y la identificación de esos motivos como los portadores evolucionados de la función biológica es un campo de investigación en rápido desarrollo.

(a) GCGGAUUUALCUCAGDDGGGAGAGCMCCAGABUGAAR.P?UGGAG7UCCUGUGTPCGAUCCACAGAAUUCGCACCA

(b) (((((((..((((........)))).(((((.......)))))).....(((((.......)))))))))))....

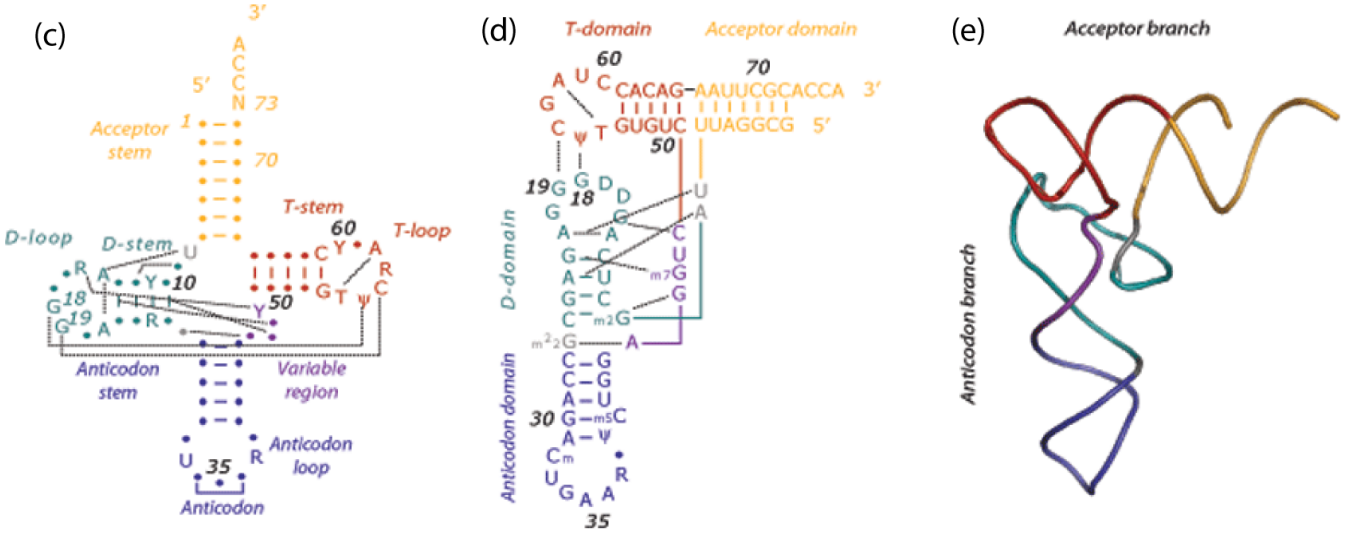

Figura 14.4: Representaciones de secuencia, estructura secundaria y terciaria de un genérico tRNA^{Phe}. (a) Secuencia. (b) Estructura secundaria en notación de puntos y paréntesis, sin interacciones terciarias. (c) La estructura secundaria dibujada en el plano, con lineas discontinuas indicando interacciones terciarias. (d) La estructura dibujada desde otro ángulo para poder ver mejor la forma real de "L" de la molecula. (e) Estructura terciaria tridimensional. Datos obtenidos de [50].

14.4. Plegamiento de RNA

14.4.1. Aspectos generales del plegamiento de RNA

El RNA de cadena sencilla puede formar enlaces consigo mismo, formando pares de bases entre nucleótidos compatibles. En general, la formación de enlaces es energéticamente favorable y de forma natural aparece una molécula con múltiples enlaces consigo mismo, representando una estructura molecular tridimensional que tiene propiedades físico-químicas distintas a la secuencia sin plegar, confiriendo funciones biológicas muy bien definidas en el entorno celular. En consecuencia, este llamado *plegamiento* de la molécula es uno de los procesos más importantes en este contexto y la predicción fiable de una estructura de RNA partiendo de su secuencia es el llamado *folding problem*. Para más información ver [60].

En el plegamiento, se habla de tres tipos de estructuras: la estructura primaria es la secuencia sin plegar (se representa por una secuencia lineal de letras con alfabeto 'ACGU'). La estructura secundaria refleja el apareamiento local de bases compatibles (Watson-Crick y *wobble*) y que siempre puede ser representado por una imagen bidimensional. Finalmente, la estructura terciaria representa la configuración espacial actual de la molécula. Frecuentemente, y sobre todo para moléculas largas, pueden estar presentes interacciones terciarias que son aquellas enlaces que van más allá de las reglas de apareamiento de bases Watson-Crick (véase Sección 14.3).

Estructura secundaria de energía libre mínima

El plegamiento de una molécula RNA es un proceso que depende de muchos factores (temperatura, concentración de iones, presencia de otras moléculas de RNA, etc.), pero el mecanismo fundamental es que nucleótidos complementarios pueden dar lugar a la formación de pares de bases. Es importante notar que para una secuencia dada, existe un gran número de estructuras compatibles diferentes, con un número de enlaces distintos y en posiciones distintas, entonces ¿cuál es la estructura real?

Por *estructuras compatibles* entendemos todas las estructuras que se pueden formar bajo las reglas de formación de pares mencionadas en la Sección 14.3 (ej. Watson-Crick plus *wobble*, bucles terminales de tamaño mínimo 3, bases que participan como mucho en un par, etc.). Pero, p. ej., un par G-C con tres enlaces de hidrógeno entre G y C implica una bajada de energía libre más grande que un par A-U o G-U. Por lo tanto, a las distintas estructuras compatibles están asociadas diferentes energías libres. En primera aproximación, la estructura secundaria observada es aquella que proporciona la bajada de energía libre total más grande, llamada *estructura MFE*, por *minimum free energy*. Por lo tanto, el problema de plegamiento es en gran medida un problema de minimización de energía.

Experimentalmente, podemos determinar las contribuciones energéticas de las distintas partes de la estructura: en general, los enlaces estabilizan, y los bucles desestabilizan la estructura. No obstante, es la formación de apilamientos la que representa la mayor parte de la bajada en energía libre en el plegamiento. Aunque permitida y teóricamente relacionada con una bajada de energía libre, la formación de pares sueltos no es muy estable en un entorno celular y poco frecuente en moléculas naturales.

Plegamiento termodinámico y cinético de la estructura secundaria

Hemos abogado por el principio de *minimización de energía libre* para explicar la estructura secundaria. Ese argumento implica que consideramos el conjunto de estructuras en el equilibrio termodinámico

(tiempo de plegamiento $t \to \infty$), y además a temperatura cero, ya que no consideramos las estructuras con mayor energía, las llamadas subóptimas. Si la temperatura es distinta a cero, la probabilidad de encontrar la estructura MFE o cualquier otra se puede calcular a través de la función de partición que sigue una distribución de Boltzmann. Además, en realidad, el plegamiento no dura infinitamente, y por lo tanto el *estado real* de la molécula plegada no tiene por que ser idéntico ni al de mínima energía ni a uno de los estados cercanos. De hecho, en muchos casos, las estructuras encontradas en secuencias naturales, no corresponden al estado de mínima energía. Eso no sólo es debido a factores externos (ej. presencia de otras moléculas) sino al proceso cinético del plegamiento mismo. En la Figura 14.5 comparamos las diferentes nociones de estructura secundaria. Vemos que la estructura MFE S_0 y la siguiente estructura en energía S_1 pertenecen a dos cuencas distintas de estructuras, separadas por una barrera alta de energía libre. Por lo tanto la estructura observada puede ser la S_1 si la secuencia al plegarse se queda atrapada en la cuenca correspondiente.

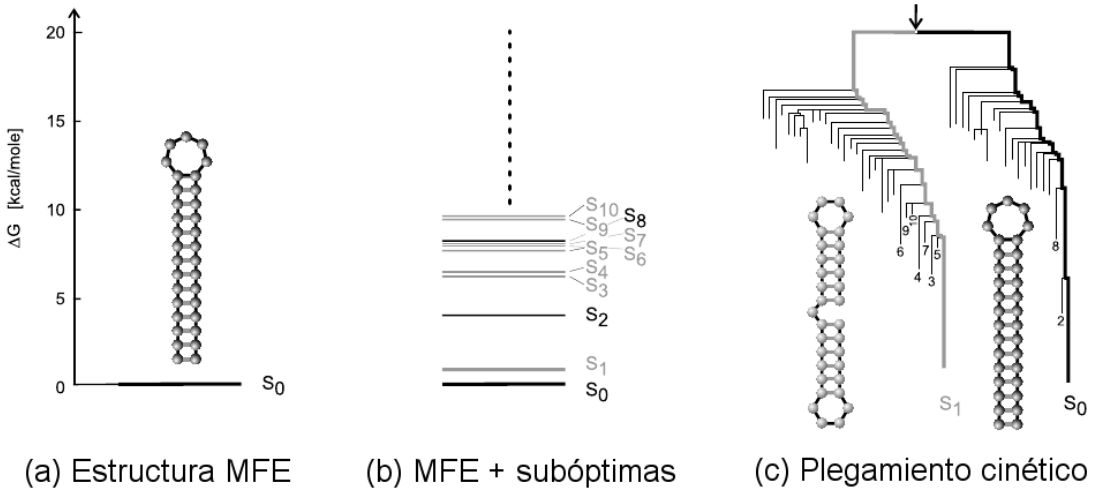

Figura 14.5: Comparación entre la estructura MFE (a), el conjunto de todas las estructuras, incluyendo las subóptimas (b) y el conjunto desde el punto de vista cinético (c). Figura basada en la Fig. 15 de [60].

Plegamiento de estructura terciaria

La formación de pares de bases tipo Watson-Crick y por lo tanto la formación de la estructura secundaria es un proceso rápido que ocurre en una escala de milisegundos. Las interacciones que dan lugar a los motivos estructurales terciarios (Sección 14.3) actúan a una escala de tiempo más lenta. Hay evidencia experimental de que la estructura secundaria del RNA se forma independientemente de la estructura terciaria [7]. Con estos datos y con datos de resonancia magnética [6], se cree que *el RNA se pliega de forma jerárquica* [15], si bien es cierta que existen excepciones [77, 78].

Por lo tanto, la estructura secundaria es un estado intermedio y forma una subestructura de la estructura final. En general, una gran parte de la energía libre de la estructura terciara final corresponde a la estructura secundaria. Por lo tanto, estudiar el plegamiento de secuencia a estructura secundaria es el primer paso a entender el proceso global.

14.4.2. Predicción de estructuras de RNA

Desde el punto de vista bioinformático, la estructura del RNA se divide en tres niveles: primaria, secundaria y terciaria (ver Sección 14.3). Como la estructura primaria es la secuencia misma, el problema de predicción de estructuras de RNA se refiere a la predicción de estructuras secundarias y terciarias, resultantes del plegamiento de la secuencia (Subsección 14.4.1).

La estructura secundaria del RNA se refiere a la colección de pares de bases de los que se compone. Por motivos de complejidad computacional, se requiere que la estructura secundaria sea planar (ver Figura 14.3(b)) y que cada nucleótido sólo pueda formar parte de un par de bases.

Formalmente, una *estructura secundaria* S sobre una secuencia de RNA $s_1, ..., s_n$ se define como un conjunto ordenado de pares que se corresponde con las posiciones de pares de bases, y que satisface las siguientes restricciones:

1. Watson-Crick o GU *wobble* pares: Si (i, j) pertenece a S, el par de bases (s_i, s_j) tiene que ser uno de los siguientes pares de bases canónicos: (A,U), (U,A), (G,C), (C,G), (G,U), (U,G).
2. Umbral (tamaño mínimo de un bucle): Si (i, j) pertenece a S, entonces $j - i > \theta$ (donde θ toma el valor 3).
3. Prohibición de pseudonudos: Si (i, j) y (k, l) pertencen a S, no se puede dar el caso en que $i < k < j < l$.
4. Prohibición de tripletes: Si (i, j) y (i, k) pertenecen a S, entonces $j == k$.

La *estructura terciaria* se refiere a la estructura tridimensional de una molécula de RNA, es decir, a las posiciones de cada uno de sus nucleótidos en tres dimensiones (o a una colección de coordenadas tridimensionales) y engloba todas las interacciones terciarias (Sección 14.3 y Subsección 14.4.1).

Es importante mencionar que la mayoría de algoritmos de predicción de estructura de RNA se centran en las posiciones de sus nucleótidos, aunque en algunos casos se encuentran algoritmos que tratan de predecir la posición de todos sus átomos.

Estructuras secundarias

La predicción de estructura secundaria de RNA es uno de los problemas más recalcitrantes de la biología computacional de RNA. Muchos algoritmos basados en distintos conceptos computacionales y/o físicos han sido desarrollados para tratar de resolverlo. Esta sección está dedicada a explicar los más utilizados a lo largo de la historia.

En el apartado anterior hemos definido la estructura secundaria, en términos generales, como una colección de pares de bases, sujeta a una serie de restricciones. Siguiendo esta definición, nos encontramos con el primer algoritmo relevante, desarrollado por *Nussinov* [48] en los anos 70. Este algoritmo pionero trata simplemente de *maximizar el número de pares de bases sin violar ninguna de las restricciones.* Se trata de un algoritmo de programación dinámica, definido por las siguientes recursiones:

$$D_{i,j} = max \begin{cases} max_{i<k<j} D_{i,k} + D_{k+1,j} \\ D_{i+1,j-1} + w_{i,j} \\ D_{i+1,j} \\ D_{i,j-1} \end{cases} \tag{14.1}$$

donde $w_{i,j} = 1$ si las posiciones i y j corresponden a los pares de bases canónicos (A,U), (U,A), (G,C), (C,G), (G,U), (U,G), o 0 en caso contrario.

Este algoritmo fue posteriormente refinado y es ahora conocido como el *algoritmo Nussinov-Jacobson* [47]. Las recursiones simplificadas son las siguientes:

$$D_{i,j} = max \begin{cases} D_{i,k} + D_{k+1,j} & donde\ i \leq k \leq j \\ D_{i+1,j-1} + w_{i,j}, \end{cases}$$

con la siguiente inicialización:

$$D_{i,j} = max \begin{cases} D_{i,i} = 0 & \forall i = 1..n \\ D_{i,i-1} = 0 & \forall i = 2..n \end{cases}$$

donde n es la longitud de la secuencia. Este algoritmo no genera necesariamente la estructura más estable, y, de hecho, su eficiencia cuando se compara con estructuras reales de RNA es muy baja. Sin embargo, es importante explicarlo ya que supuso un gran paso adelante en el área de predicción de estructura secundaria.

El siguiente grupo de algoritmos está basado en conceptos termodinámicos, más concretamente en *buscar la estructura con mínima energía libre*. El primer concepto es el *modelo del vecino más cercano* (*Nearest Neighbor* en inglés). En vez de atribuir energías a pares de bases, asigna energías a 'pares de pares' de bases. Las contribuciones de las combinaciones distintas han sido determinadas mediante técnicas experimentales [45, 76].

En un siguiente nivel, nos encontramos con una serie de motivos estructurales que aportan una energía libre a la molécula completa: bucle terminal, *bulge*, bucle interior y bucle múltiple (comparar con la Figura 14.3(c)). Dicha energía libre (G°) está compuesta por dos términos: *entalpía* (H°) y *entropía* (S°), relacionados por la siguiente ecuación[2]:

$$G^\circ = H^\circ - TS^\circ \tag{14.2}$$

donde T es la temperatura.

Así pues, nos encontramos en la situación en la que queremos calcular la estructura secundaria que minimiza la energía libre. El primer algoritmo para predecir la estructura de mínima energía libre fue desarrollado por *Zuker* [80]. Al igual que en el caso anterior, se trata de un algoritmo de programación dinámica definido por una serie de recursiones. Estas recursiones, son, sin embargo, un tanto más complejas, y requieren una introducción más detallada.

Por lo tanto, es más adecuado explicar las recursiones para calcular la función de partición. La *función de partición* es un concepto de física estadística que describe las propiedades de un sistema en equilibrio termodinámico. Es sabido que un sistema en equilibrio termodinámico con una reserva de temperatura tiene probabilidades p de ocupar un estado con energía E multiplicado por su correspondiente factor de Boltzmann. Sea $\mathcal{X} = \{X_1, \cdots, X_n\}$ un sistema de estados, donde el estado X_i tiene una energía E_i. El sistema sigue una *distribución de Boltzmann* con temperatura T si y solo si

$$Pr[X_i] = exp(-\beta E_i)/Z \tag{14.3}$$

donde $Z = \sum_i exp(-\beta E_i)$ y $\beta = (RT)^{-1}$.

[2]Simplificando la ecuación correcta: $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$.

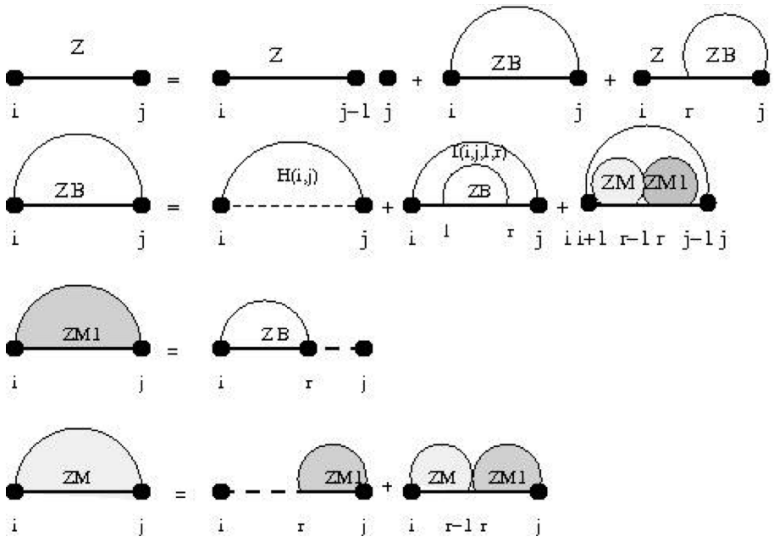

Figura 14.6: Representación gráfica de las cuatro matrices que incorpora el algoritmo de Zuker.

De esta forma podemos definir la función de partición Z de una molécula de RNA (en función de la temperatura):

$$Z = \sum_{P \in \Omega} exp(-E(P)/RT) \tag{14.4}$$

donde:

- P es una estructura determinada,
- Ω es el espacio de estructuras,
- $E(P)$ es la energía de la estructura,
- R es la constante de gases,
- T es la temperatura absoluta en grados Kelvin,
- $exp(-E(P)/RT)$ se conoce como el factor de Boltzmann.

Así pues, dada una secuencia de RNA $a_1, \ldots, a_n$, para todo $1 \leq i \leq j \leq n$, la función de partición $Z_{i,j}$ se define como $\sum_S e^{-E(S)/RT}$, donde la suma se calcula sobre todas las posibles estructuras secundarias S de $a[i,j]$, $E(S)$ es la energía libre mínima de la estructura S.

Ahora estamos preparados para presentar las recursiones correspondientes al algoritmo de Zuker [3] para calcular la función de partición de una molécula de RNA. Dicho algoritmo se basa en el uso de cuatro matrices, cuyas definiciones son (ver también la Figura 14.6):

- $Z_{i,j}$: función de partición sobre todas las estructuras secundarias en $a[i,j]$.
- $ZB_{i,j}$: función de partición sobre todas las estructuras secundarias en $a[i,j]$, que contienen el par de bases (i,j).
- $ZM_{i,j}$: función de partición sobre todas las estructuras secundarias en $a[i,j]$, sujetas a la restricción de que $a[i,j]$ es parte de un multiloop y tiene *como mínimo* un componente.

[3] En realidad, las recursiones para la función de partición reciben el nombre de recursiones de McCaskill [46], si bien el algoritmo para predecir la estructura de mínima energía libre es conocido como el algoritmo de Zuker [80].

- $ZM1_{i,j}$: función de partición sobre todas las estructuras secundarias en $a[i,j]$, sujetas a la restricción de que $a[i,j]$ es parte de un multiloop y tiene *exactamente* un componente. Además, se *requiere* que i forme un par de bases en el intervalo $[i,j]$; es decir, (i,r) es un par de bases, para algún $i < r \leq j$.

y cuyas recursiones son las que siguen:

$$Z_{i,j} = \begin{cases} 1 & \text{si } j-i \leq \theta \\ Z_{i,j-1} + ZB_{i,j} + \sum_{k=i+1}^{j-\theta-1} Z_{i,k-1} \cdot ZB_{k,j} & \text{en caso contrario} \end{cases}$$

$$ZM1_{i,j} = \begin{cases} 0 & \text{si } j-i \leq \theta \\ \sum_{k=i+\theta+1}^{j} \exp\left(-\frac{c\cdot(j-k)}{RT}\right) \cdot ZB_{i,k} & \text{en caso contrario} \end{cases}$$

$$ZM_{i,j} = \begin{cases} 0 & \text{si } i \leq j \text{ y } j-i \leq \theta \\ \sum_{k=i}^{j-\theta-1} \exp(-\frac{b+c\cdot(k-i)}{RT}) ZM1_{k,j} + & \\ \sum_{k=i}^{j-\theta-2} \exp\left(-\frac{b}{RT}\right) \cdot ZM_{i,k} \cdot ZM1_{k+1,j} & \text{en caso contrario} \end{cases}$$

$$ZB_{i,j} = \begin{cases} 0 & \text{si } j-i \leq \theta \\ Z_{i,j}(S) + Z_{i,j}(H) + Z_{i,j}(B) + Z_{i,j}(I) + Z_{i,j}(M) & \text{en caso contrario} \end{cases}$$

donde:

$$\begin{aligned} Z_{i,j}(S) &= \exp\left(-\frac{E(i,i+1,;j-1,j,T)}{RT}\right) \cdot ZB_{i+1,jj-1} \\ Z_{i,j}(H) &= \exp\left(-\frac{H(j-i-1,T)}{RT}\right) \\ Z_{i,j}(LB) &= \sum_{k=i+3}^{j-\theta-2} \exp\left(-\frac{B(k-i-1,T)}{RT}\right) \cdot ZB_{k,j-1} \\ Z_{i,j}(RB) &= \sum_{k=i+\theta+2}^{j-3} \exp\left(-\frac{B(j-k-1,T)}{RT}\right) \cdot ZB_{i+1,k} \\ Z_{i,j}(I) &= \sum_{\ell-i-2\leq 2}^{j-\theta-3} \sum_{j-r-1+\ell-i-2\leq 30}^{j-2} \exp\left(-\frac{I((\ell-i-1)+(j-r-1))}{RT}\right) \cdot ZB_{\ell,r} \\ Z_{i,j}(M) &= \exp\left(-\frac{a+2b}{RT}\right) \cdot \sum_{k=i+\theta+3}^{j-\theta-2} ZM_{i+1,k-1} \cdot ZM1_{k,j-1} \end{aligned}$$

donde:

- $E(i,i+1;j,j-1,T)$ es la energía correspondiente al apilamiento de los pares de bases (i,j) y $(i+1,j-1)$ a temperatura T,
- $H(l,T)$ es la energía de un hairpin de longitud l a temperatura T,
- $B(l,T)$ es la energía de un bulge de longitud l a temperatura T,

- $I(l,T)$ es la energía de un bucle interno de longitud l a temperatura T,
- a es una penalización energética por empezar un bucle múltiple,
- b es una penalización energética por el numero de componentes de un bucle múltiple.

Teniendo en cuenta estas recursiones para el cálculo de la función de partición de una molécula de RNA, podemos inferir fácilmente las recursiones para calcular la estructura MFE: basta con cambiar los máximos por sumas y las sumas por multiplicaciones, y reemplazar los factores de Boltzmann por energías.

También es posible calcular la probabilidad de que dos nucleótidos formen un par de bases siguiendo el *algoritmo de McCaskill* [46]. Este algoritmo computa las probabilidades de pares de bases siguiendo la fórmula:

$$p(i,j) = \frac{\sum -\{S : (i,j) \in S\} exp(-E(S)/RT)}{\sum_S exp(-E(S)/RT)}$$

Es decir, los probabilidades de pares de bases de la distribución de Boltzmann, donde el numerador comprende la suma de todas las estructuras secundarias en las que ocurre el par de bases (i,j), y en el denominador todas las posibles estructuras secundarias de RNA dada su secuencia.

Este tipo de algoritmo es posiblemente el más extendido, conocido y utilizado. Actualmente existen 3 implementaciones distintas que se diferencian en detalles como las energías libres o el tratamiento de *dangles* (nucleótidos sin par al final de una hélice) y apilamiento coaxial: `UNAfold` de Zucker [44], `RNAfold` de Hofacker [31] (dentro del software Vienna Package) y `RNAstructure` de Mathews [56].

Finalmente, discutiremos otro enfoque importante dentro del área de la predicción de estructura secundaria de RNA, la llamada *estructura de máxima esperanza de precisión* (o *Maximum Expected Accurate* en inglés). El concepto de estructura de máxima esperanza de precisión fue introducido por Do [21] y su definición es la siguiente: dada un secuencia $a_1, \cdots, a_n$ de RNA, la estructura de máxima esperanza de precisión S es aquella que maximiza la siguiente formula:

$$\sum_{(i,j)\in S} 2\alpha \cdot p(i,j) + \sum_{i\, desapareado} \beta q_i$$

donde

$$p(i,j) = \text{probabilidad del par de bases } (i,j)$$

y

$$q_i = 1 - \sum_{i<j} p(i,j) - \sum_{j<i} p(j,i).$$

La estructura de máxima esperanza de precisión puede ser calculada con un algoritmo de programación dinámica similar al de Nussinov-Jacobson con probabilidades obtenidas a través de una gramática estocástica de contexto libre como en [40], o con probabilidades obtenidas usando el algoritmo de McCaskill como en [42].

Pseudonudos

La predicción de RNA pseudonudos representa un paso entre la predicción de estructuras secundarias con una complejidad relativamente baja y la predicción de estructuras terciarias con una complejidad alta. A nivel práctico, también interacciones del tipo *kissing loop* pueden ser considerados como pseudonudos. Se puede demostrar que el problema de encontrar pseudonudos arbitrarios es NP-duro. Por lo tanto hay que reducir la complejidad del problema.

Podemos distinguir entre algoritmos heurísticos y recursivos. Además, tanto programas de predicción de estructuras secundarias como terciarias pueden incluir pseudonudos. Recomendamos estudiar las programas de predicción antes de utilizarlos, se puede consultar una lista actualizada de programas en Wikipedia[4]

Empezamos con las *recursiones*. Un método muy eficaz se basa en el *algoritmo de Reeder y Giegerich* que escala como $O(n^4)$ en tiempo y $O(n^2)$ en espacio (Programa `PKnotsRG` [51]). Utiliza una recursión del tipo Zuker y se limita a los pseudonudos más simples, con un sólo posible entrelazamiento entre dos *stacks* y sin *bulges*. A pesar de estas restricciones, ese algoritmo incluye al tipo de pseudonudos más frecuente y estable y es un método rápido.

Otro algoritmo recursivo es el de Rivas y Eddy [57]. Incluye la posibilidad de *kissing hairpin* y *three-knot* y escala como $O(n^6)$ en tiempo y $O(n^4)$ en espacio. Mencionamos los algoritmos de Akutsu [3] y de Chen, Condon y Jabbari [13] como métodos con complejidad y prestaciones intermedias.

Presentamos dos métodos de predicción de pseudonudos basado en heurísticas. Estos algoritmos son de interés porque son rápidos y pueden alcanzar una buena precisión [11]. Uno se llama `ILM` (*Iterated Loop Matching*) [58] y utiliza una iteración tipo Nussinov como plegamiento herárquico. Otro método es `HotKnots` [55], basado en elegir elementos de la estructura secundaria con energía libre relativamente baja, para sucesivamente añadir elementos hacía la estructura final.

Otro algoritmo heurístico es `Probknot` [8]. En este caso, en vez de usar la energía libre se basa en la estructura de máxima esperanza de precisión. Usando las probabilidades de dos nucleótidos de formar un par de base, la estructura de máxima esperanza de precisión puede ser extendida a un problema de apareamiento máximo, donde los nodos del grafo son nucleótidos y las aristas tienen un peso igual a la probabilidad. En este contexto, resolver el problema de apareamiento máximo resulta en un conjunto de pares de bases que puede contenter pseudonudos. Aunque el problema de apareamiento máximo se puede resolver en tiempo polinómico, Probknot usa un algoritmo más rápido y más eficiente.

Estructuras terciarias

En la literatura, es posible encontrar métodos que hablan de estructura terciaria de RNA, pero en realidad sólo tienen en cuenta estructuras secundarias con pseudonudos o con tripletes de bases y pares de bases no-canónicos. Este es el caso de [35] donde se presenta un algoritmo capaz de encontrar la estructura secundaria de mínima energía libre con tripletes y pares de base no-canónicos. Para ello, utiliza unas complejas recursiones que derivan del concepto de *2-diagrams* sin cruces.

Sin embargo, existe algún método para calcular la estructura terciaria de una molécula de RNA. El más conocido pertenece a las llamadas `MC-tools` de Major [49]. `MC-fold` primero encuentra estructuras secundarias que maximicen la frecuencia de cuartetos de nucleótidos que aparecen en las bases de datos de estructuras reales de RNA. Estos cuartetos pueden ser o no ser pares de bases apilados. Esta estructura secundaria es transformada en terciaria gracias a `MC-Sym`. `MC-Sym` tiene en cuenta todos los átomos y se basa en especificar los llamados *dihedral angles* que definen las posiciones del esqueleto azúcar-fosfato y de las bases. Estos ángulos estan restringidos a sus valores más frecuentes, calculados a partir de estructuras reales.

Muy recientemente encontramos un nuevo método para el cálculo de la estructura terciaria de RNA en [54]. Este método también utiliza `MC-Sym`, aunque sustituye `MC-Fold` por `RNA-MoIP`. `RNA-MoIP`

[4] List of RNA structure prediction software. `http://en.wikipedia.org/wiki/List_of_RNA_structure_prediction_software`

refina estructuras secundarias utilizando programación entera para acomodar estructuras terciarias, en muchas ocasiones, reemplazando pares de bases canónicos por no-canónicos.

Finalmente, cabe reseñar el reciente esfuerzo por impulsar el diseño de algoritmos de predicción de estructura terciaria de RNA fomentado por la competición conocida como RNA-Puzzles[5].

14.4.3. Otros problemas relacionados con el plegamiento de RNA

Hay muchos problemas y herramientas relacionadas con la estructura del RNA. Dejando de lado el área de las herramientas para la búsqueda de RNA no codificante, describiremos a continuación los ejemplos más relevantes.

Hibridización de RNA

Existen varios algoritmos para determinar la estructura secundaria de mínima energía de hibridización de RNA (o de RNA con DNA). Este problema tiene aplicaciones en el estudio de *micro-arrays* y en el estudio de ciertos procesos biológicos como el sistema CRISPR en bacterias o el proceso de splicing de RNA con exclusividad mutua de casetes como en el caso del gen Dscam en *Drosophila* [30].

Para calcular la mínima energía libre de hibridización se usa un algoritmo similar al de Zucker, en el cual las dos cadenas se juntan y se tiene un tratamiento especial de las estructuras en torno a la posición de juntura. Hay varias herramientas que resuelven este problema, como `RNAcofold` del Vienna Package o `RNAhybrid` dentro de `UNAfold`. También existen otras herramientas como `RNAduplex` o `RNAup` (ambas dentro del Vienna Package). `RNAduplex` calcula todas las posibles estructuras locales de hibridización entre dos cadenas de RNA sin considerar estructuras intra-moleculares, y `RNAup` calcula la mejor estructura local de hibridización teniendo en cuenta que la cadena de mayor longitud forma una estructura secundaria que hay que romper antes de formarse la hibridización.

RNAbor

`RNAbor` es una herramienta desarrollado por el Clote Lab [26] en la que se pueden calcular la probabilidad de encontrar estructuras que difieren en k pares de bases con respecto a una estructura inicial. Dicha estructura inicial puede ser tanto la estructura de mínima energía libre como la estructura vacía (ningún par de bases) o cualquier otra estructura aleatoria. Esto se consigue usando un algoritmo de programación dinámica similar al de McCaskill en el cual se va concentrando la probabilidad de cada valor k simultáneamente.

Tanto `RNAbor` como su reciente versión `FFTbor` [63] en la que se mejora la velocidad usando la transformada de Fourier pueden utilizarse para detectar riboswitches, para estudiar el espacio de plegamiento o incluso para aproximar la velocidad de plegamiento de RNA.

RNAmutants

`RNAmutants` es otra herramienta desarrollada por el Clote Lab [75]. De una forma similar al caso anterior, `RNAmutants` calcula la probabilidad (y la estructura de mínima energía libre) para todas las

[5] RNA-Puzzles. `http://paradise-ibmc.u-strasbg.fr/rnapuzzles`

secuencias que difieren en k nucleótidos (k nucleótidos que cambian de valor, no que se añaden o se eliminan) de la secuencia original.

`RNAmutants` puede ser utilizado para estudiar el espacio mutacional de una secuencia de RNA, identificando posiciones más débiles o más robustas ante posibles mutaciones. `RNAmutants` también sirve para comparar la robustez de diferentes familias de RNA.

Mapeo secuencia a estructura secundaria

Si queremos conocer para una secuencia dada el conjunto de estructuras compatibles, también queremos saber cual el conjunto total de estructuras para todas las secuencias. Existe un nivel de degeneración entre la secuencia y la estructura secundaria. Mientras hay 4^n secuencias de longitud n, hay un número más bajo de estructuras secundarias, $S(n)$. Grüner el al. derivaron aproximaciones para ese número, $S(n) \approx 0{,}7131 \times n^{-3/2}(2{,}2888)^n$, si se permite la formación de pares sueltos, y $S(n) \approx 1{,}4848 \times n^{-3/2}(1{,}849)^n$, si se prohíbe la formación de pares sueltos [32].

Destacamos varias propiedades importantes del mapeo entre secuencia y estructura secundaria: (a) Degeneración: como hay más secuencias que estructuras, sobre todo para n grande, hay secuencias distintas que pliegan en la misma estructura. (b) Distribución de estructuras: existen pocas estructuras que son el resultado del plegamiento de muchas secuencias (estructuras frecuentes) mientras hay muchas estructuras que son estructuras raras. Es una distribución ancha con una cola larga. (c) Estructuras frecuentes están distribuidos de forma homogénea en el espacio de las secuencias: de una secuencia aleatoria en pocas mutaciones se puede obtener una secuencia que pliega en una estructura frecuente. (d) Redes neutrales: Las secuencias que pliegan en estructuras frecuentes forman una red extensa en el espacio de secuencias de tal manera que si la longitud de la molécula es suficientemente larga, esa red neutral percola el espacio de secuencias.

En la misma línea del cálculo del número de estructuras secundarias, existe todo un campo a lo que se ha venido a llamar Combinatoria de RNA. Dentro de este campo destacamos: cálculo de el número de estructuras secundarias canónicas y saturadas [16], cálculo de la distancia entre el 5' y el 3' de una molécula de RNA [17], etc.

Comparación de estructuras

Mientras la diferencia entre secuencias se puede cuantificar con la *distancia Hamming* (número de nucleótidos distintos), para comparar estructuras existen una variedad de métodos distintos. Aunque se puede definir una distancia Hamming entre estructuras basado en el alfabeto punto-paréntesis, no se considera una medida muy apta para estructuras, y se prefiere la *distancia base-pair* que es el número de pares que uno tiene que abrir y cerrar para convertir una estructura en otra. Además, existe la posibilidad de definir distancias basado en la formulación de árbol para una estructura, y se define una distancia a través de cambios que uno tiene que hacer para convertir una estructura en otra. Un ejemplo es la *distancia tree-edit* que junto con las otras medidas están presentadas en [60].

Un concepto similar es el de agrupar estructuras según su similitud, reduciendo la descripción de la estructura secundaria. Una posible definición es el *Shape* que es una representación de estructura secundaria en la que las hélices se colapsan, por ejemplo, (((...((...))..(((...)))..).)) pasa a ser [[ ] [ ]] [41]. Otra posibilidad es la definición de familias de estructuras como en [66, 69].

14.4.4. Biología Sintética de RNA: plegamiento inverso

La *biología sintética* es una disciplina emergente con ramificaciones que van desde la detección de moléculas dentro de la célula a la creación de genomas sintéticos y nuevas formas de vida. Grupos pioneros en este campo han obtenido resultados espectaculares como, por ejemplo, la síntesis combinatoria de redes de genes, la síntesis de genomas usando *BioBricks*, y la reacción de hibridización en cadena, en la cual monómeros estables de DNA se juntan sólo ante la exposición a un fragmento de DNA objetivo. La mayor parte de los trabajos en el campo de la biología sintética se concentran en lo que se puede considerar como genómica sintética, como por ejemplo, la regulación sintética de genes [14] y el desarrollo de bloques genéticos, gracias a los cuales, se pueden construir nuevos genomas [64].

Sin embargo, esta sección versa sobre la biología sintética de RNA. Como hemos visto en los apartados anteriores, y al igual que en el caso de las proteínas, la estructura de RNA determina en gran medida su función. Dada la gran cantidad de nuevos tipos de RNA con funciones complejas, y dada la actual variedad de algoritmos para predicción de estructura de RNA, de alineamiento estructural, de búsqueda de motivos estructurales, etc; parece claro que algunos de los más importantes avances en la biología sintética tratarán el diseño y validación experimental de nuevas estructuras sintéticas de RNA.

Plegamiento inverso de RNA

Dado que el cálculo de la estructura secundaria de mínima energía libre puede ser realizado de forma eficiente y que la determinación de la estructura de RNA con pseudonudos y la estructura terciaria son NP-completos [43], nos centraremos en el problema del plegamiento inverso en cuanto se refiere a la estructura secundaria.

Como la estructura secundaria de RNA es una parte esencial de la estructura terciaria y el plegamiento tiene lugar de forma jerárquica (ver Cap. 14.4), cualquier solución al problema del plegamiento inverso para estructura secundaria es, claramente, un gran paso hacia el diseño de RNA funcional.

Dada una estructura secundaria S sobre una cadena desconocida de longitud n, el *problema del plegamiento inverso* consiste en encontrar la secuencia de RNA $a_1, \cdots, a_n$ (es decir, la palabra de longitud n del alfabeto $\sigma = A, C, G, U$) cuya estructura mínima de energía libre es S.

Aunque no ha sido demostrado formalmente, se cree que el problema del plegamiento inverso de RNA es NP-duro.

Existen varios algoritmos para resolver el problema del plegamiento inverso de RNA: `RNAinverse` [34], `RNA-SSD` [5], `INFO-RNA` [9], `MODENA` [71], `NUPACK-DESIGN` [79], `Inv` [28]. Todos estos algoritmos pueden ser clasificados como heurísticos: comienzan con una secuencia inicial que es mejorada de forma iterativa hasta encontrar una solución o terminar por algún otro criterio como tiempo límite o máximo número de iteraciones.

El primer algoritmo que se encuentra en la literature es `RNAinverse`, el cual forma parte del Vienna Package [31, 34]. `RNAinverse` divide la estructura objetivo S en subunidades más pequeñas e intenta encontrar la secuencia de RNA usando un camino adaptativo o algoritmo avaro. Las posiciones de la secuencia inicial son mutadas, estas mutaciones se aceptan si la función objetivo mejora. En este caso, la función objetivo es la distancia de Hamming entre la estructura de mínima energía libre de la secuencia y la estructura objetivo S. `RNAinverse` puede devolver la solución correcta, una solución aproximada o ninguna solución, dependiendo de la dificultad de la instancia.

`RNA-SSD` [5] es un algoritmo distinto y muy eficienciente, si bien comparte la misma filosofía de divide y vencerás, dividiendo la estructura de forma jerárquica en subunidades. En comparación con

RNAinverse, RNA-SSD utiliza una inicialización más sofisticada e implementa un algoritmo de búsqueda local estocástica, en vez de un simple camino adaptativo. RNA-SSD es capaz de encontrar la secuencia correcta para estructuras de más de mil nucleótidos de longitud.

El tercer algoritmo es INFO-RNA [9]. La principal diferencia con los anteriores algoritmos reside en su inicialización, la cual se basa en un algoritmo de programación dinámica para elegir la secuencia $s_1, \ldots, s_n$ que es compatible con S y tiene la menor energía libre. Aunque la energía libre $E(s_1, \ldots, s_n; S)$ de la estructura objective S en $s_1, \ldots, s_n$ es menor o igual a la energía libre $E(s'_1, \ldots, s'_n; S)$ para todas las secuencias $s'_1, \ldots, s'_n$ que son compatibles con S, esto no significa que la estructura de mínima energía libre of $s_1, \ldots, s_n$ es la estructura objetivo S. INFO-RNA es, por lo menos, igual de eficiente que RNA-SSD, y debido a su especial inicialización, devuelve secuencias de RNA cuya energía libre es muy baja (con alto contenido de GCs). Aunque esto pueda parecer beneficioso, estas secuencias de alto contenido de GCs suele tener poco parecido con secuencias reales de RNA.

El cuarto, MODENA [71], es bastante diferente a todos los anterior. MODENA usa el conocido algoritmo genético NSGA2 [19] para encontrar el conjunto débil de Pareto de soluciones óptimas respecto a dos funciones objetivo: estabilidad de la estructura (mínima energía libre) y similitud (distancia entre la estructura de mínima energía libre y la estructura objetivo S).

NUPACK-DESIGN [79], es el punto de partida de un pionero proyecto del laboratorio de Pierce, para diseñar moléculas de RNA que más tarde son validadas tanto *in vitro* como *in vivo*. NUPACK-DESIGN utiliza un enfoque parecido a RNA-SSD, aunque en este caso, NUPACK-DESIGN intenta encontrar secuencias con mínimo *defecto de colectividad* [20].

Dada una secuencia de RNA $s = s_1, \ldots, s_n$ con respecto a una estructura objetivo S, el *defecto de colectividad* se define como el valor esperado de nucléotidos cuya estado de emparejamiento difiere del que tienen en la estructura objetivo S. Formalmente tenemos:

$$n(s, S_0) = n - \sum_{1 \le i,j \le n} p^*_{i,j} \cdot I[(i,j) \in S] - \sum_{1 \le i \le n} p^*_{i,n+1} \cdot I[i \text{ desapareado en } S]$$

donde I es la función indicatriz y, para cada posición fija $1 \le i \le n$, se define la función de probabilidad $p^*_{i,j}$, para todo j en $[1, n+1]$, simetrizando p para valores $1 \le i, j \le n$, y así pues definiendo $p^*_{i,n+1} = 1 - \sum_{j>i} p_{i,j} - \sum_{j<i} p_{j,i}$.

Finalmente, el algoritmo Inv [28] utiliza una rutina de búsqueda local estocástica para encontrar la secuencia cuya mínima energía libre con psuedonudos es la *3-noncrossing* estructura objetivo. Una estructura *3-noncrossing* es una estructura (posiblemente con psuedonudos) en la cual no es posible encontrar 3 (o más) pares de bases que se cruzan mutuamente. Inv es un algoritmo de programación dinámica de tiempo exponencial [37] y utiliza el hecho de que cada *3-noncrossing* estructura tiene una única descomposición en hélices.

En contraposición a todos los algoritmos anteriores, un nuevo enfoque utiliza programación con restricciones y es completo. RNAifold [29] esta implementado en COMET [72] y tiene la capacidad de encontrar todas las secuencias cuya estructura de mínima energía libre es la estructura objetivo.

14.5. Evolución de RNA

14.5.1. El RNA como modelo evolutivo

La evolución molecular cubre una enorme área de investigación. Esta comprende desde la química prebiótica y las preguntas sobre el origen de la vida, hasta el diseño artificial de moléculas y la selección y

evolución in vitro con sus aplicaciones en nano y biotecnología, pasando por muchos aspectos relacionados con el origen y las relaciones entre las especies, el estudio de la evolución viral y bacteriana y sus implicaciones médicas. En este texto, no pretendemos dar una visión completa de todo el campo, sino centrarnos en los enfoques teóricos del uso de las poblaciones de moléculas de RNA como un modelo para entender la evolución de replicadores simples. Y para ello, comenzamos identificando las dos propiedades que todo sistema evolutivo debe tener: (i) un *mecanismo que introduzca variabilidad genética* en la población (en nuestro caso es la mutación puntual de nucleótidos) y (ii) una *diferenciación de los genomas con respecto a su capacidad replicativa* (*fitness*) que permita la selección de los más aptos (mediante una presión selectiva).

Las poblaciones de *moléculas de RNA* son un modelo muy adecuado para estudiar procesos evolutivos porque *incorporan, en una única entidad molecular, tanto genotipo como fenotipo* [25]. Mientras que los errores en el proceso de replicación introducen mutaciones en la secuencia de RNA (genotipo), la selección actúa sobre la función (fenotipo) de la molécula. Como en muchos casos la estructura espacial de la molécula es crucial para su función bioquímica, la estructura de una molécula de RNA puede considerarse como una representación del fenotipo.

No obstante, es muy difícil obtener la estructura real (es decir, tridimensional y con todos sus detalles) en la que pliega una secuencia arbitraria (ver Sección 7.7), y en la práctica conocemos con alta precisión sólo las estructuras de algunas moléculas fundamentales en la biología actual (como el tRNA, de 76 nucleótidos). Sin embargo, la estructura secundaria de una secuencia sí que se puede conocer con más precisión y también predecir con fidelidad mediante algoritmos informáticos. Por esta razón se suele utilizar *la estructura secundaria como representación mínima del fenotipo*, y por extensión, de la función bioquímica de la molécula. Como consecuencia de todo esto, conviene recordar que existen varios niveles de degeneración en el sistema: muchas secuencias pueden plegar en la misma estructura, y diferentes estructuras pueden tener la misma función bioquímica.

14.5.2. Redes neutrales de RNA: concepto y propiedades básicas

La idea de la *evolución neutral* fue introducida por Kimura [39] con el fin de explicar un resultado por entonces muy sorprendente: *un gran número de mutaciones observadas en proteínas, DNA, o RNA, no tienen efecto sobre el fitness de dichas moléculas*. La neutralidad es particularmente importante en la evolución de cuasiespecies [22], poblaciones de replicadores de alta tasa de mutación que están formadas por un gran número de fenotipos diferentes -y muchísimos más genotipos- donde la alta diversidad y la exploración constante del espacio de genomas se convierte en una estrategia adaptativa. Algunos ejemplos relevantes de cuasiespecies son los virus de RNA (VIH, Ebola, Gripe, etc...) y los replicadores de alta tasa de mutación en el contexto del mundo de RNA prebiótico.

Como ya se ha comentado, el RNA es un modelo ideal para el estudio de la evolución molecular, y de hecho las secuencias de RNA plegando en sus estructuras secundarias de energía libre mínima son probablemente el modelo más utilizado en la literatura de la relación genotipo-fenotipo [61]. *Cada estructura secundaria de RNA*, utilizada como aproximación del fenotipo o función biológica de la molécula, *tiene asociada una red neutral de genotipos*, donde los nodos representan todas las secuencias que pliegan en dicha estructura. Dos nodos de la red están conectados por un enlace cuando sus secuencias están a *distancia de Hamming* 1, es decir, cuando difieren en un solo nucleótido [52] (véase la Figura 14.7). Para una longitud dada l de la secuencia de RNA, tenemos por lo tanto tantas redes neutrales como estructuras secundarias distintas (o fenotipos distintos). Conviene resaltar el hecho de que cada red neutral puede contener uno o más componentes aislados entre sí, es decir, que no siempre mediante mutaciones puntuales una secuencia puede recorrer la totalidad de secuencias que pliegan en la misma estructura secundaria. En este último caso, la red neutral se compone de *subredes*.

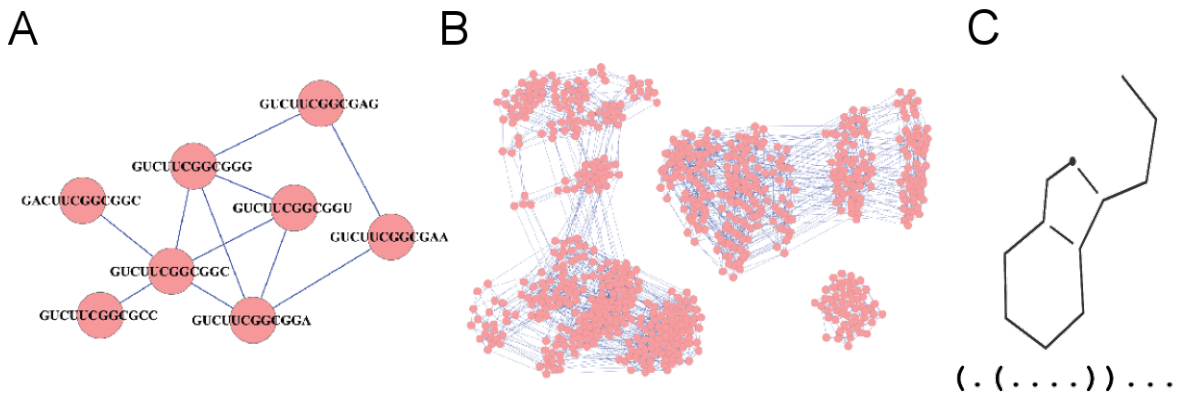

Figura 14.7: Construcción de una red neutral asociada a una estructura secundaria de RNA dada. (A) Esquema de la construcción: Los nodos de la red son todas las secuencias que pliegan en la misma estructura secundaria, y se conectan entre sí si están a una distancia de Hamming uno, es decir, si difieren sólo en un nucleótido. (B) Red neutral asociada a la estructura secundaria de longitud 12 (.(....))..., mostrada en (C). Nótese que, aunque todas las secuencias de una red neutral pliegan en la misma estructura secundaria, la red neutral puede estar dividida. En este caso, existen 3 subredes aisladas de tamaños 404, 341 y 55. Figura modificada de [2].

Puesto que el RNA consta de cadenas de 4 nucleótidos distintos (cuyas bases nitrogenadas son A, G, C y U), las redes neutrales asociadas a una longitud de secuencia l son *subredes inmersas* en la macrorred formada por todas las secuencias posibles de longitud l. Esta macrorred es regular, tiene tamaño 4^l, dimensión l, y cada nodo tiene grado $3l$, porque cada nucleótido puede sufrir tres mutaciones distintas, y cada secuencia tiene l nucleótidos. (Véase la Sección 19.2 para una introducción básica a la teoría de redes complejas.)

Los estudios analíticos del tamaño de las redes neutrales, es decir, de la cantidad de secuencias de longitud l compatible con una estructura secundaria dada (lo que se conoce como degeneración genotipo-fenotipo), han revelado que *las redes neutrales de genotipos son astronómicamente grandes incluso para valores moderados de la longitud de la secuencia*. Sólo por citar un par de ejemplos, diremos que existen aproximadamente 10^{28} secuencias compatibles con la estructura de una molécula de tRNA (que tiene una longitud $l = 76$), mientras que las estructuras funcionales más cortas conocidas actualmente, que tienen una longitud l de unos 12 nucléotidos, pueden obtenerse a partir de más de 10^6 secuencias diferentes [2]. De hecho, una aproximación válida para secuencias suficientemente largas es que existen $S_l = 1{,}4848 \times l^{-3/2}(1{,}8488)^l$ estructuras secundarias para secuencias de longitud l [62], y por lo tanto el tamaño medio de una red neutral crece con l como $4^l/S_l = 0{,}673 \times l^{3/2}2{,}1636^l$, cantidad que crece enormemente con l. Este valor medio no es sin embargo representativo de la distribución real de los tamaños de redes neutrales, que es una función muy ancha sin una media bien definida y con una cola también muy ancha [62, 66]. De hecho, el espacio de secuencias de RNA de longitud l está dominado por un número relativamente pequeño de estructuras comunes que son muy abundantes y que se encuentran habitualmente en la naturaleza como motivos estructurales de moléculas de RNA funcional [23, 27]. Las redes neutrales correspondientes a dichas estructuras comunes percolan el espacio de secuencias [33, 53] y facilitan así la exploración de un gran número de estructuras alternativas. Esto es posible porque *las diferentes redes neutrales están profundamente entrelazadas*: a partir de cualquier secuencia aleatoria, y mediante pocos pasos sobre la red (o mutaciones), es posible llegar a todas las estructuras comunes [53].

14.5.3. Modelización matemática de la evolución sobre redes neutrales de RNA

Ecuaciones dinámicas

Más adelante en este capítulo, en la Subsección 14.5.5, nos centraremos en el caso general de la evolución de poblaciones de RNA sobre un paisaje de *fitness* rugoso que tiene en cuenta de forma diferenciada todas las estructuras posibles de RNA. En él, una población puede acceder a través de mutaciones a todo el espacio de secuencias de tamaño 4^l en su camino hacia una *estructura objetivo*. Sin embargo, comenzaremos suponiendo que, de todas las estructuras secundarias posibles en las que pueden plegar todas las secuencias de RNA de longitud l, sólo una es funcional (su función específica dependerá del contexto, y es irrelevante ahora). Por lo tanto, el *fitness* de las secuencias de una única red es máximo, mientras que el *fitness* de todas las secuencias pertenecientes a otras redes neutrales es 0. Con esta hipótesis, la evolución de una población de RNA a través del espacio de genotipos debido a mutaciones se limita a la red neutral funcional (o a cada subred de la misma si está partida en subredes), ya que suponemos que sólo la progenie que se mantiene dentro de la red neutral sobrevive. Veamos la forma de modelizar este proceso haciendo uso de la teoría de cadenas de Markov y matrices de transición [1].

Cada nodo i en la red, que representa a un genotipo distinto, tiene un número $n_i(t)$ de secuencias en un tiempo dado t. Hay $i = 1, \ldots, m$ nodos en la red, cada uno con un grado, o número de vecinos, k_i. La población total es $N = \sum_i n_i(t)$, y suponemos $N \to \infty$ para evitar efectos estocásticos debido al tamaño finito de la población. La distribución inicial de las secuencias en la red a $t = 0$ es $n_i(0)$. Las secuencias de longitud l, formadas por los 4 nucleótidos diferentes de que consta el RNA, tienen a lo sumo $3l$ vecinos. Llamamos $\{v_i\}$ al conjunto de los k_i vecinos del nodo i. En cada paso de tiempo, las secuencias en cada nodo se replican. Las nuevas secuencias mutan a uno de los $3l$ vecinos más cercanos con probabilidad μ, y se mantienen iguales a la secuencia madre con una probabilidad $1 - \mu$. En nuestro caso, $0 < \mu \leq 1$. El caso singular $\mu = 0$ se excluye para evitar una dinámica trivial y garantizar una evolución hacia un estado de equilibrio único. Por lo tanto, con probabilidad $k_i/(3l)$ las secuencias mutadas existen en la red neutral y en ese caso se suman a la población de los nodos vecinos correspondientes. De lo contrario, caen fuera de la red y desaparecen.

Las *ecuaciones que describen la dinámica de la población sobre la red neutral* presentada en el párrafo anterior son:

$$n_i(t+1) = (2-\mu)n_i(t) + \frac{\mu}{3l} \sum_{j=\{v_i\}} n_j(t) \tag{14.5}$$

y de forma matricial

$$\vec{n}(t+1) = (2-\mu)I\vec{n}(t) + \frac{\mu}{3l} A\vec{n}(t) \tag{14.6}$$

donde I es la matriz identidad y A es la matriz de adyacencia de la red, cuyos elementos son $A_{ij} = 1$, si los nodos i y j están conectados y $A_{ij} = 0$ en caso contrario (véase la Sección 19.2).

La dinámica asociada al sistema cuya evolución la define una matriz de transición M viene dada por $\vec{n}(t+1) = M\vec{n}(t)$. En nuestro caso particular, la matriz de transición M se define como

$$M = (2-\mu)I + \frac{\mu}{3l} A \tag{14.7}$$

Estudio del estado asintótico de la población

Llamemos $\{\lambda_i\}$ al conjunto de autovalores de M, con $\lambda_i \geq \lambda_{i+1}$, y $\{\vec{u}_i\}$ los autovectores correspondientes (para informaciím general sobre autovalores y autovectores, véase cualquier libro de álgebra lineal, por ejemplo [12]). Los autovectores verifican que $\vec{u}_i \cdot \vec{u}_j = 0$, $\forall i \neq j$ y $|\vec{u}_i| = 1$, $\forall i$. Una matriz es irreducible cuando el grafo correspondiente es conexo, y en nuestro caso cualquier par de nodos i y j de la red están conectados a través de mutaciones por definición. Si además de irreducible, una matriz cumple la condición $M_{ii} > 0, \forall i$, entonces se dice que la matriz M es primitiva. La matriz de transición M presentada en la Ecuación 14.7 es efectivamente primitiva, y el teorema de Perron-Frobenius asegura que, en el intervalo de valores de μ utilizados, el autovalor mayor de M es positivo, es mayor que el resto de autovalores, y su autovector asociado también es positivo (es decir,$(\vec{u}_1)_i > 0$, $\forall\ i$). La verificación de este teorema es importante porque nos permite dar los pasos mostrados a continuación (véase [36], por ejemplo, para profundizar en el teorema de Perron-Frobenius).

La dinámica asociada al sistema, plasmada en la Ecuación 14.6, puede ser escrita de la forma:

$$\vec{n}(t) = M^t \vec{n}(0) = \sum_{i=1}^{m} \lambda_i^t \alpha_i \vec{u}_i \tag{14.8}$$

donde α_i es la proyección de la condición inicial sobre el autovector i de M, $\alpha_i = \vec{n}(0)\, \vec{u}_i$.

Además, como $\lambda_1 > |\lambda_i|$ por el teorema de Perron-Frobenius, $\forall i > 1$, el estado asintótico de la población es proporcional al autovector que corresponde al mayor autovalor, $\vec{u}_1$:

$$\lim_{t\to\infty} \left(\frac{\vec{n}(t)}{\lambda_1^t \alpha_1} \right) = \vec{u}_1 \tag{14.9}$$

mientras que el mayor autovalor λ_1 nos da la tasa de crecimiento de la población en el equilibrio (en ausencia de reescalamiento). Si se normaliza la población $\vec{n}(t)$ tal que $|\vec{n}(t)| = 1$ después de cada generación, se cumple que $\vec{n}(t) \to \vec{u}_1$ cuando $t \to \infty$. Estos resultados son importantes, porque afirman que, independientemente de la condición inicial, la población de secuencias tiende asintóticamente hacia una distribución constante sobre la red que es igual al primer autovector de la matriz de transición.

Tiempo al estado asintótico

Otra cantidad dinámica relevante y con directas implicaciones biológicas es el tiempo que la población tarda en alcanzar dicho equilibrio. Para obtenerlo, partimos de la Ecuación 14.8, que describe la dinámica transitoria hacia el equilibrio a partir de la condición inicial $\vec{n}(0)$. La distancia $\Delta(t)$ al estado de equilibrio se puede escribir como:

$$\Delta(t) \equiv \left| \frac{M^t \vec{n}(0)}{\lambda_1^t \alpha_1} - \vec{u}_1 \right| = \left| \sum_{i=2}^{m} \vec{\Delta}_i(t) \right| = \left| \sum_{i=2}^{m} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^t \vec{u}_i \right| \tag{14.10}$$

A fin de estimar cuántas generaciones transcurren antes de que se alcance el equilibrio, fijamos una tolerancia ϵ, y definimos el *tiempo al equilibrio* t_ϵ como el número de generaciones necesarias para que se cumpla que $\Delta(t_\epsilon) < \epsilon$. En general, t_ϵ se puede aproximar a primer orden por:

$$t_\epsilon^1 \simeq \frac{\ln |\alpha_2/\alpha_1| - \ln \epsilon}{\ln \lambda_1/\lambda_2} \tag{14.11}$$

y en definitiva:

$$t_\epsilon \propto \left(\ln \frac{\lambda_1}{\lambda_2}\right)^{-1} \tag{14.12}$$

Esta aproximación resulta ser extraordinariamente acertada en la mayoría de los casos gracias a la supresión exponencialmente rápida de las contribuciones debidas a los términos de orden superior (puesto que $\lambda_i \geq \lambda_{i+1}$, $\forall i$). Se puede perder precisión, sin embargo, cuando $\lambda_3 \approx \lambda_2$, cuando la condición inicial de $\vec{n}(0)$ es tal que $\alpha_3 \gg \alpha_2$, o cuando ϵ es tan grande que la población está lejos del equilibrio y $\Delta(t)$ se sigue rigiendo por λ_3 y los autovalores de orden superior.

Influencia de la topología de red en la dinámica

Llamemos $\{\gamma_i\}$ al conjunto de autovalores de la matriz de adyacencia A, $\gamma_i \geq \gamma_{i+1}$, y $\{\vec{w}_i\}$ el conjunto de autovectores correspondientes. De la Ecuación 14.7, obtenemos:

$$M\vec{w}_i = (2-\mu)I\vec{w}_i + \frac{\mu}{3l}A\vec{w}_i = \left[(2-\mu) + \frac{\mu}{3l}\gamma_i\right]\vec{w}_i \tag{14.13}$$

Por lo tanto, los autovectores de la matriz de adyacencia A son también autovectores de la matriz de transición M, $\vec{u}_i \equiv \vec{w}_i$, $\forall i$, lo que demuestra que el estado asintótico de la población sólo depende de la topología de la red neutral [73]. Los autovalores de ambas matrices están relacionados a través de:

$$\lambda_i = (2-\mu) + \frac{\mu}{3l}\gamma_i \tag{14.14}$$

donde el conjunto $\{\gamma_i\}$ no depende de la tasa de mutación μ. En resumen, la matriz de adyacencia contiene toda la información de los estados finales, mientras que la matriz de transición nos da información cuantitativa sobre la dinámica hacia el equilibrio.

Como se ve en la Ecuación 14.14, el valor mínimo de λ_1 se obtiene en el límite de una población evolucionando a una tasa de mutación muy alta ($\mu \to 1$) y para moléculas muy grandes ($l \to \infty$). En este límite, todos los autovalores de M se vuelven independientes de la topología precisa de la red y $\lambda_i \to 1$, $\forall i$. En este caso extremo todas las secuencias hijas caen fuera de la red y mueren, pero la población se mantiene constante a través de la población original.

Obtención de un resultado biológico relevante: la tendencia a la robustez mutacional

El *grado medio de la población* $K(t)$ en el tiempo t se define como:

$$K(t) = \frac{\vec{k}\cdot\vec{n}(t)}{\sum_i n_i(t)} \tag{14.15}$$

donde $\vec{k}$ es el vector grado y tiene como componentes el grado de los $i = 1, \ldots, m$ nodos de la red. En el límite de $t \to \infty$, se obtiene el grado medio en el equilibrio:

$$K(t\to\infty) = K = \frac{\vec{k}\cdot\vec{u}_1}{\sum_i (u_1)_i} \tag{14.16}$$

Se definen como k_{min}, k_{max} y $\langle k \rangle = \sum_i k_i/m$ el grado mínimo, máximo, y medio de la red, respectivamente. Un cálculo muy simple (basado en la identidad entre los autovectores de la matriz de adyacencia A y la de transición M nos dice que el grado medio de la población en el equilibrio, K, es igual al autovalor más grande γ_1 de la matriz de adyacencia, también conocido como *radio espectral* de la red [73].

Además, el teorema de Perron-Frobenius para grafos conectados, con pesos no negativos y simétricos, establece límites en el grado medio $\langle k \rangle$: Cuando $k_{min} < k_{max}$, es decir, en la medida que el grafo no es regular, se cumple:

$$k_{min} < \langle k \rangle < \gamma_1 = K < k_{max} \, . \tag{14.17}$$

Por lo tanto, el grado medio K de la población en el equilibrio será mayor que el grado medio $\langle k \rangle$ de la red, lo que indica que la población selecciona regiones con conectividad por encima del promedio de la red. Las implicaciones evolutivas de este resultado son importantes. El grado de cada nodo i refleja su robustez ante mutaciones, porque cuanto mayor sea dicho grado más probabilidades tendrá la secuencia de RNA de, en caso de mutar, mantener la estructura y en definitiva la función. Por lo tanto, la relación $K > \langle k \rangle$ nos dice que una población de secuencias de RNA que evolucione sobre una red neutral tiende de forma natural a la robustez ante mutaciones, y en consecuencia a la fijación de la estructura (Véase la Figura 14.8).

14.5.4. Limitaciones y líneas futuras de la modelización de la evolución sobre redes neutrales de RNA

Los enormes tamaños de las redes neutrales de RNA imposibilitan los estudios sistemáticos de este tipo de redes para RNAs de longitudes que no sean muy pequeñas, y suponen un reto importante a la bioinformática. Actualmente, sólo se han hecho estudios exhaustivos plegando la totalidad de las secuencias de RNA de longitud por debajo de unos 20 nucleótidos, donde el número de secuencias a plegar no supera los 10^{12}. Por ejemplo, recientemente se estudiaron las propiedades topológicas de todas las redes neutrales asociadas a $l = 12$ [2]. En la Tabla 14.2 se muestra una comparativa entre las propiedades topológicas de dichas redes neutrales y las de las redes de topología más habitual, las aleatorias Erdös-Renyi (ER) y las libres de escala Barabási-Albert (BA). Como queda claro en la tabla, la topología de las redes neutrales dista mucho de ser aleatoria, y de hecho el grado medio de la red $\langle k \rangle$, y por lo tanto la robustez ante mutaciones, crece con el tamaño de la red. Este hecho refuerza la hipótesis de que las estructuras que se ven en la naturaleza son aquellas cuyas redes neutrales son más grandes. Al argumento de que al ser más extensas, una secuencia que evoluciona aleatoriamente por el espacio de genomas encontrará dicha red más fácilmente, se une el hecho de que, una vez hallada, la mayor robustez a las mutaciones de la red mantendrá con más fidelidad a la población de secuencias dentro de sus márgenes.

Sin embargo, no hay seguridad de que los resultados obtenidos para moléculas de RNA de longitudes pequeñas sean extensibles a moléculas más largas, como puede ser el tRNA ($l = 76$), y no digamos ya a los virus de RNA, cuyo material genético tiene del orden de miles de nucleótidos, o incluso más. El desafío computacional, teórico y experimental es fabuloso, y sin duda este problema requiere un enfoque distinto al utilizado hasta ahora. Además, no podemos olvidar que para desarrollar los cálculos mostrados en este apartado, hemos supuesto que existe una única red neutral funcional, mientras que el resto han sido obviadas. En la naturaleza, en muchas ocasiones varias estructuras secundarias distintas, aunque similares, pueden desarrollar la misma función, aunque su eficiencia sea distinta. Cada una de las redes neutrales asociadas a estas estructuras puede estar conectada en una red de redes, donde las

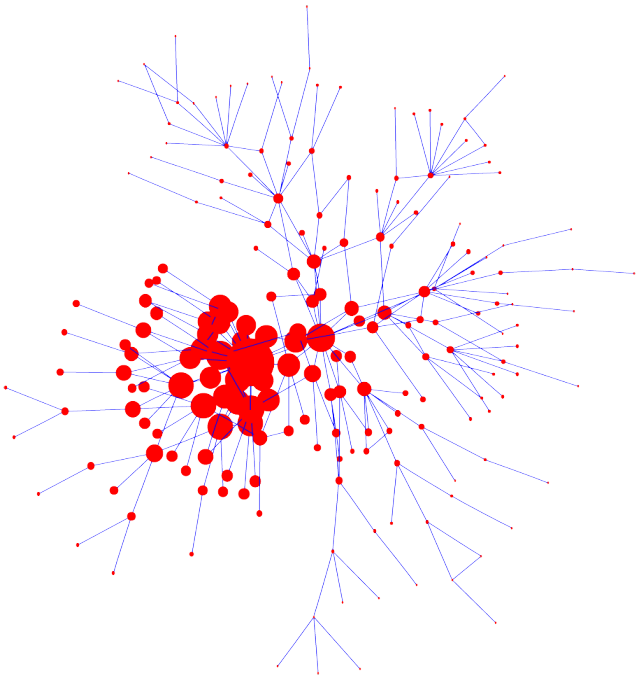

Figura 14.8: Tendencia a la robustez mutacional. Independientemente de la distribución inicial, una población de replicadores que evoluciona sobre una red compleja tiende en el equilibrio a las zonas más conectadas. En la figura hemos hecho evolucionar una población siguiendo la Ecuación 14.6 sobre una red neutral artificial. El tamaño de cada nodo es proporcional a la población alcanzada en el equilibrio, y muestra con claridad la tendencia natural a poblar la región con mayor grado medio de la red, y por lo tanto al refuerzo de las secuencias más robustas ante mutaciones. Figura modificada de [1].

	Redes neutrales	**Aleatorias (ER)**	**Libres de escala (BA)**
$p(k)$	pico único	distr. de Poisson	ley de potencias ($\sim k^{-3}$)
$\langle k \rangle(N)$	$\sim \ln N$	constante	constante
$C(k)$	$\sim k^{-1}$	constante ($\frac{\langle k \rangle}{N}$)	constante ($\sim N^{-0,75}$)
$C(N)$	$\sim (\ln N)^{-1}$	$\sim N^{-1}$	$\sim N^{-0,75}$
$\langle d \rangle(N)$	$\sim \ln N$	$\sim \ln N / \ln \langle k \rangle$	$\sim \ln N / \ln \ln N$
$k_{nn}(k)$	$\sim k^{0,75}$	constante ($\langle k^2 \rangle / \langle k \rangle$)	no trivial [4]
Asortatividad	asort. ($r > 0$)	no asort. ($r \to 0$)	no asort. ($r \to 0$)

Tabla 14.2: Comparación de la topología de las redes neutrales de longitud de secuencia $l = 12$ con las redes aleatorias (Erdös-Renyi) y libres de escala (Barabási-Albert). Las cantidades reflejadas son la distribución de grado $p(k)$, el clustering en función del grado $C(k)$ y en función del tamaño de la red $C(N)$ respectivamente, el camino medio $\langle d \rangle(N)$, el grado medio de los vecinos $k_{nn}(k)$, y la asortatividad r. (Véase la Sección 19.2 para más información acerca de estas cantidades.)

secuencias van mutando y van saltando de unas redes a otras y variando en consecuencia su *fitness*. Si cada una de estas redes es representada por un único nodo, entonces obtenemos lo que se conoce como *redes de fenotipos*, campo de estudio abierto recientemente y que promete aportar importantes

resultados a esta línea de investigación [18].

14.5.5. Evolución dirigida a una estructura objetivo y definición de distancia estructural

En los apartados anteriores hemos supuesto que una población de moléculas de RNA de longitud l se mueve estrictamente encima de una red neutral y que las moléculas que pliegan en otra estructura secundaria tienen *fitness* cero. Este *paisaje de fitness*, conocido como de *pico único* (*single-peak*), es obviamente una aproximación muy simplificada. En esta sección consideramos una generalización representando un paisaje de *fitness* rugoso que tiene en cuenta de forma diferenciada todas las estructuras posibles, y donde las poblaciones evolucionan hacia una estructura objetivo.

Asumimos que una población puede acceder a través de mutaciones a todo el espacio de secuencias de tamaño 4^l y que conocemos sus estructuras secundarias asociadas. Elegimos una estructura como estructura objetivo, es decir, con fitness óptimo. Ahora, las demás estructuras no tienen fitness zero, sino un *fitness* que depende de la similitud con la estructura objetivo. Para comparar estructuras entre sí, se han introducido diferentes medidas de distancia: *base-pair*, *Hamming*, y *tree-edit* (véase Sección 14.4.3). Se puede demostrar que esas medidas establecen una métrica y que podemos utilizarlas para definir una función de *fitness*. Una posibilidad es que la probabilidad de replicación de una molécula esté dada por $p \propto exp(-\beta d)$, donde d es la distancia entre la estructura de la molécula y la estructura objetivo y β la magnitud de la presión selectiva.

Como el mapeo de secuencia a estructura es altamente complejo, hay que recurrir a métodos numéricos y estadísticos para estudiar el proceso evolutivo de una población. El esquema es el siguiente:

1. Elegir una estructura objetivo. Puede ser cualquier estructura, pero obviamente los casos más relevantes son aquellos en los que la elegida representa una estructura biológicamente relevante, p.ej. una horquilla (*hairpin*), una cabeza de martillo (*hammerhead*), tRNA, o similar. El espacio de secuencias, estructuras, la imprecisión del plegamiento y los recursos computacionales crecen fuertemente con la longitud de la molécula, por lo que muchos estudios están enfocados a moléculas cortas, de orden 10^1 a 10^2 nucleótidos.

2. Establecer el tamaño N de la población.

3. Formar una población inicial, que puede constar de secuencias al azar o elegidas expresamente.

4. Determinar las estructuras secundarias de las secuencias. Se pueden utilizar programas de libre acceso como el RNAfold del Vienna Package [34], véase Sección 14.4.2.

5. Determinar las distancias de las estructuras a la estructura objetivo.

6. Replicar la población con el método Wright-Fisher, teniendo en cuenta las probabilidades normalizadas de replicación:

$$p(d_i) = \frac{\exp(-\beta d_i)}{\sum_{i=1}^{N} \exp(-\beta d_i)}. \tag{14.18}$$

   Entre las posibilidades para elegir la magnitud β, destacamos dos particularmente relevantes: en función de la longitud de la molécula, p.ej., $\beta = 1/l$, o en función de una propiedad variable, p.ej., $\beta = 1/d$, siendo $d = \sum_{i=1}^{N} d_i/N$ la distancia media de la población, y reflejando así una selección dependiente de la frecuencia.

7. Al crear una nueva generación de la población (que reemplaza a la anterior), hay que incluir una fuente de variabilidad, en el caso más simple mutaciones puntuales con una tasa de mutación por nucleótido μ.

El proceso evolutivo avanza de generación en generación al repetir los pasos (4)-(7).

En el caso general, en la población inicial no hay secuencias que pliegan en la estructura objetivo, y por lo tanto la primera fase de la evolución está marcada por la búsqueda de dicha estructura objetivo. Cuando la población ya la ha encontrado por primera vez, empieza la segunda fase, en la que la estructura se fija y el número de moléculas con la estructura correcta crece en la población. Los tiempos de busqueda dependen no sólo de la tasa de mutación, sino también de la estructura elegida [69]. Si la tasa de mutación es demasiado alta, es posible que la estructura objetivo vuelva a perderse. Una vez fijada la estructura, la población puede llegar a su estado asintótico, caracterizado por una distancia media mínima y una fracción máxima de moléculas con estructura objetivo. Como la mutación sigue operando, ese estado asintótico es estacionario sólo con respecto a cantidades medias. Además, si N es demasiado bajo, pueden darse efectos de tamaño finito.

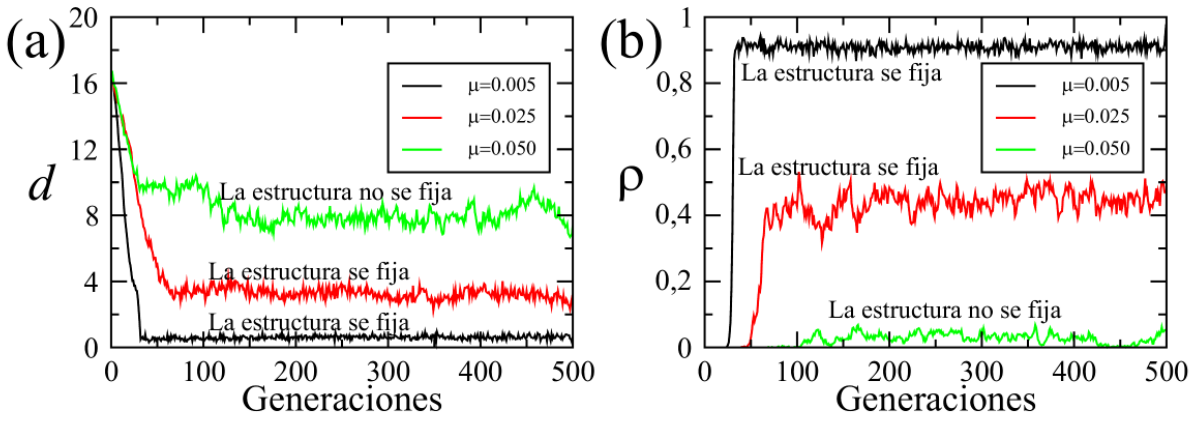

Figura 14.9: Comportamiento evolutivo típico de una población de RNA para tres tasas de mutación distintas. Mostramos d, la distancia base-pair media a la estructura objetivo de la población, y ρ, la densidad de estructuras correctas en la población. El tamaño de la población es 602, $\beta = 1$, y la estructura objetivo es una horquilla (o *hairpin*) de longitud 35 nucleótidos.

En la Figura 14.9, mostramos un ejemplo numérico de lo que acabamos de exponer. En ella se observa el comportamiento de una población que evoluciona para tres valores distintos de la tasa de mutación. Las variables son d, la distancia base-pair media a la estructura objetivo de la población, y ρ, la densidad de estructuras correctas en la población. Si la tasa de mutación es pequeña, la población termina conteniendo un alto número de estructuras correctas, y la distancia media es baja. Si aumenta la tasa de mutación, se encuentran un número menor de estructuras correctas, y dicha distancia media aumenta. Si la tasa de mutación supera un cierto límite, la estructura objetivo no se mantiene de forma permanente en la población (ρ tiende a cero), y decimos que la estructura no se ha *fijado* en la población.

Desde que se implantó por primera vez un sistema de evolución de una población de RNA utilizando su estructura secundaria [24, 38], este modelo ha sido modificado muchas veces, aunque manteniendo su filosofía original. La función de selección puede ser exponencial, pero se han utilizado otras funciones

(lineales y no-lineales). Además, se han utilizado distintos tipos de distancia RNAEstructura: mientras la *base-pair* es la que más se asemeja a un proceso biofísico (el abrir y cerrar de enlaces), no puede ser utilizada si la población contiene moléculas de distintos tamaños, en cuyo caso se recurre a la distancia *tree-edit*. Muchas propiedades de las poblaciones en evolución sólo dependen cuantativamente, no cualitativamente, de estas variables.

Obviamente, la función de selección puede ser ampliada para no sólo tener en cuenta la distancia a la estructura objetivo, sino también la distancia a una secuencia (p.ej. un trozo de una secuencia conservada) si la hubiera, u otro criterio de optimización, p.ej. una minimización de energía [67]. Finalmente, el marco del modelo también permite que el tamaño de la población no sea constante, para poder abarcar los llamados cuellos de botella [68].

En resumen, una población de moléculas de RNA en evolución permite estudiar procesos diversos: los tiempos de búsqueda que dependen de la tasa de mutación y el tipo de estructura, la adaptación de poblaciones en función de la fuerza de selección y mutación, o la robustez respecto a cambios de entorno (estructuras objetivo distintas, tamaños de población), entre otros muchos.

14.6. Bibliografía

[1] J. Aguirre, J. Buldú, and S. Manrubia. Evolutionary dynamics on networks of selectively neutral genotypes: Effects of topology and sequence stability. *Phys. Rev. E*, 80:066112, 2009.

[2] J. Aguirre, J. M. Buldú, M. Stich, and S. C. Manrubia. Topological structure of the space of phenotypes: The case of RNA neutral networks. *PLoS ONE*, 6:e26324, 2011.

[3] T. Akutsu. Dynamic programming algorithms for rna secondary structure prediction with pseudoknots. *Discr. Appl. Math.*, 104:45–62, 2000.

[4] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[5] M. Andronescu, A. Fejes, F. Hutter, H. Hoos, and A. Condon. A new algorithm for rna secondary structure design. *J. Mol. Biol.*, 336:607–624, 2004.

[6] M. Bailor, X. Sun, and H. Al-Hashimi. Topology links rna secondary structure with global conformation, dynamics, and adaptation. *Science*, 327:202–206, 2010.

[7] A. Banerjee, J. Jaeger, and D. Turner. Thermal unfolding of a group i ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32:153–163, 1993.

[8] S. Bellaousov and D. Mathews. ProbKnot: fast prediction of rna secondary structure including pseudoknots. *RNA*, 16:1870–1880, 2010.

[9] A. Busch and R. Backofen. Info-rna, a fast approach to inverse rna folding. *Bioinformatics*, 22:1823–1831, 2006.

[10] S. Butcher and A. Pyle. The molecular interactions that stabilize rna tertiary structure: Rna motifs, patterns, and networks. *Acc. Chem. Res.*, 44:1302–1311, 2011.

[11] S. Cao and S.-J. Chen. Predicting rna pseudoknot folding thermodynamics. *Nucl. Acids Res.*, 34:2634–2652, 2006.

[12] M. Castellet and I. Llerena. *Algebra lineal y geometría*. Editorial Reverté, 2009.

[13] H.-L. Chen, A. Condon, and H. Jabbari. An $O(n(5))$ algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *J. Comput. Biol.*, 16:803–815, 2009.

[14] Y. Chen, M. Jensen, and C. Smolke. Genetic control of mammalian T-cell proliferation with synthetic rna regulatory systems. *Proc. Natl. Acad. Sci. USA*, 107:8531–8536, 2010.

[15] S. Cho, D. Pincus, and D. Thirumalai. Assembly mechanisms of rna pseudoknots are determined by the stabilities of constituent secondary structures. *Proc. Natl. Acad. Sci. USA*, 106:17349–17354, 2009.

[16] P. Clote, E. Kranakis, D. Krizanc, and B. Salvy. Asymptotics of canonical and saturated rna secondary structures. *J. Bioinform. Comput. Biol.*, 7:869–893, 2009.

[17] P. Clote, Y. Ponty, and J.-M. Steyaert. Expected distance between terminal nucleotides of rna secondary structures. *J. Math. Biol.*, 2011.

[18] M. Cowperthwaite, E. Economo, W. Harcombe, E. Miller, and L. Meyers. The ascent of the abundant: how mutational networks constrain evolution. *PLoS Comp. Biol.*, 4:e1000110, 2008.

[19] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley and Sons, 2001.

[20] R. Dirks, M. Lin, E. Winfree, and N. Pierce. Paradigms for computational nucleic acid design. *Nucl. Acids Res.*, 32:1392–1403, 2004.

[21] C. Do, M. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 15:330–340, 2005.

[22] M. Eigen. Selforganization of matter and evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

[23] W. Fontana, D. Konings, P. Stadler, and P. Schuster. Statistics of rna secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[24] W. Fontana and P. Schuster. A computer-model of evolutionary optimization. *Biophys. Chem.*, 26:123–147, 1987.

[25] W. Fontana and P. Schuster. Shaping space: The possible and the attainable in rna genotype-phenotype mapping. *J. Theor. Biol.*, 194:491–515, 1998.

[26] E. Freyhult, V. Moulton, and P. Clote. Boltzmann probability of rna structural neighbors and riboswitch detection. *Bioinformatics*, 23:2054–2062, 2007.

[27] H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of rna secondary motifs using graph theory with implications for rna design. *Nucl. Acids Res.*, 31:2926–2943, 2003.

[28] J. Gao, L. Li, and C. Reidys. Inverse folding of rna pseudoknot structures. *Algorithms Mol. Biol.*, 5(27), 2010.

[29] J. Garcia-Martin, P. Clote, and I. Dotu. RNAiFold: A constraint programming algorithm for rna inverse folding and molecular design. *J. Bioinform. Comput. Biol.*, 2013. in press.

[30] B. Graveley. Mutually exclusive splicing of the insect dscam Pre-mRNA directed by competing intronic rna secondary structures. *Cell*, 123:65–73, 2005.

[31] A. Gruber, R. Lorenz, S. Bernhart, R. Neubock, and I. Hofacker. The vienna rna websuite. *Nucl. Acids Res.*, 36:70–74, 2008.

[32] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. Hofacker, P. Stadler, and P. Schuster. Analysis of rna sequence structure maps by exhaustive enumeration. i. neutral networks. *Monatsh. Chem.*, 127:355–374, 1996.

[33] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. Hofacker, P. Stadler, and P. Schuster. Analysis of rna sequence structure maps by exhaustive enumeration. ii. structures of neutral networks and shape space covering. *Monatsh. Chem.*, 127:375–389, 1996.

[34] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of rna secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[35] C. Honer zu Siederdissen, S. Bernhart, P. Stadler, and I. Hofacker. A folding algorithm for extended rna secondary structures. *Bioinformatics*, 27:129–136, 2011.

[36] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 2nd edition, 2013.

[37] F. Huang, W. Peng, and C. Reidys. Folding 3-noncrossing rna pseudoknot structures. *J. Comput. Biol.*, 16:1549–1575, 2009.

[38] M. A. Huynen, D. A. M. Konings, and P. Hogeweg. Multiple coding and the evolutionary properties of RNA secondary structure. *J. Theor. Biol.*, 165:251–267, 1993.

[39] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[40] H. Kiryu, T. Kin, and K. Asai. Robust prediction of con-sensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, 23:434–441, 2007.

[41] W. Lorenz, Y. Ponty, and P. Clote. Asymptotics of rna shapes. *J. Comput. Biol.*, 15:31–63, 2008.

[42] Z. Lu, J. Gloor, and D. Mathews. Improved rna secondary structure prediction by maximizing expected pair accuracy. *RNA*, 15:1805–1813, 2009.

[43] R. Lyngso and C. Pedersen. Rna pseudoknot prediction in energy-based models. *J. Comput. Biol.*, 7:409–427, 2000.

[44] N. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.

[45] D. Mathews, J. Sabina, M. Zuker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of rna secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[46] J. McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[47] R. Nussinov and A. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proc. Natl. Acad. Sci. USA*, 77:6309–6313, 1980.

[48] R. Nussinov, G. Pieczenik, J. Griggs, and D. Kleitman. Algorithms for loop matchings. *SIAM J. Appl. Math.*, 35:68–82, 1978.

[49] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers rna structure from sequence data. *Nature*, 452:51–55, 2008.

[50] J. Putz, B. Dupuis, M. Sissler, and C. Florentz. Mamit-trna, a database of mammalian mitochondrial trna primary and secondary structures. *RNA*, 13(8):1184–90, 2007.

[51] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:104, 2004.

[52] C. Reidys, C. Forst, and P. Schuster. Replication and mutation on neutral networks. *Bull. Math. Biol.*, 63:57, 2001.

[53] C. Reidys, P. Stadler, and P. Schuster. Generic properties of combinatory maps - neutral networks of rna secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[54] V. Reinharz, F. Major, and J. Waldispühl. Towards 3D structure prediction of large rna molecules: an integer programming framework to insert local 3D motifs in rna secondary structure. *Bioinformatics*, 28:207–214, 2012.

[55] J. Ren, B. Rastegari, A. Condon, and H. Hoos. Hotknots: Heuristic prediction of rna secondary structures including pseudoknots. *RNA*, 11:1494–1504, 2005.

[56] J. Reuter and D. Mathews. RNAstructure: software for rna secondary structure prediction and analysis. *BMC Bioinf.*, 11:129, 2010.

[57] E. Rivas and S. Eddy. A dynamic programming algorithm for rna structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[58] J. Ruan, G. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of rna secondary structures with pseudoknots. *Bioinformatics*, 20:58–66, 2004.

[59] T. Schlick. *Molecular Modeling and Simulation: An interdisciplinary Guide*. Springer, 2010.

[60] P. Schuster. Prediction of rna secondary structures: from theory to models and real molecules. *Rep. Prog. Phys.*, 69:1419–1477, 2006.

[61] P. Schuster. Prediction of RNA secondary structures: from theory to models and real molecules. *Rep. Prog. Phys.*, 69:1419, 2006.

[62] P. Schuster, W. Fontana, P. Stadler, and I. Hofacker. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc. Roy. Soc. (London) B*, 255:279–284, 1994.

[63] E. Senter, S. Sheik, I. Dotu, Y. Ponty, and P. Clote. Using the fast fourier transform to accelerate the computational search for rna conformational switches. *PLoS ONE*, 2012. in press.

[64] R. Shetty, D. Endy, and T. Knight Jr. Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.*, 2:5, 2008.

[65] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of trna sequences and sequences of trna genes. *Nucl. Acids Res.*, 26:148–153, 1998.

[66] M. Stich, C. Briones, and S. Manrubia. On the structural repertoire of pools of short, random rna sequences. *J. Theor. Biol.*, 252:750–763, 2008.

[67] M. Stich, E. Lázaro, and S. Manrubia. Phenotypic effect of mutations in evolving populations of rna molecules. *BMC Evol. Biol.*, 10:46, 2010.

[68] M. Stich, E. Lázaro, and S. Manrubia. Variable mutation rates as an adaptive strategy in replicator populations. *PLoS ONE*, 5:e11186, 2010.

[69] M. Stich and S. Manrubia. Motif frequency and evolutionary search times in rna populations. *J. Theor. Biol.*, 280:117–126, 2011.

[70] J. Stombaugh, C. Zirbel, E. Westhof, and N. Leontis. Frequency and isostericity of rna base pairs. *Nucl. Acids Res.*, 37:2294–2312, 2009.

[71] A. Taneda. Modena: a multi-objective rna inverse folding. *Adv. Appl. Bioinf. Chem.*, 4(1), 2011.

[72] P. Van Hentenryck and L. Michel. *Constraint-Based Local Search*. MIT Press, 2005.

[73] E. vanNimwegen, J. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*, 96:9716–9720, 1999.

[74] G. Varani and W. McClain. The gu wobble base pair. *EMBO Rep.*, 1:18–23, 2000.